

LREC 2026

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) @ LREC 2026

Workshop Proceedings

Editors
Antonios Anasastasopoulos, Stella Markantonatou, Angela Ralli, Marcos Zampieri, Stavros Bompolas, Vivian Stamou

16 May 2026

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) @ LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-60-9

Preface

The proceedings contains the 34 papers accepted for presentation at the Workshop on Dialects in NLP — A Resource Perspective (DialRes-LREC2026, https://dialres.github.io/dialres/index.html), co-located with the Fifteenth Biennial Language Resources and Evaluation Conference (LREC). The conference was held at the Palau de Congressos de Palma in Palma de Mallorca, Spain, from 11 to 16 May 2026, with the DialRes workshop taking place on 16 May 2026.

For this first edition of DialRes, we received 43 submissions. We are deeply grateful to the dialect studies and NLP communities for embracing DialRes as a venue for presenting their work. This strong response highlights the growing interest in dialectal research and confirms the need for a dedicated forum focusing on the development and use of dialectal resources.

The accepted papers cover both living and historical dialects from nearly all continents. Contributions on historical dialects include studies on English, German, Heptanesian Greek, Old Irish, and Transcarpathian varieties. Research on living dialects spans a wide range of linguistic contexts, including Arabic, Aromanian, Bangla, Basque, Formosan languages, German, Italian, Kurdish, Modern Greek, and Slavic varieties (such as Pomak and Ukrainian), as well as Wancho.

The contributions address a broad spectrum of topics, ranging from the collection and development of oral and written dialectal resources—such as corpora, treebanks, benchmarks, and specialized databases—to their application in dialectometry and dialect classification. Many papers employ state-of-the-art methodologies, including Automatic Speech Recognition (ASR), neural parsing techniques, and Large Language Models (LLMs).

We would also like to express our gratitude to the Archimedes Research Unit of the Athena Research Center and the UniDive COST Action for endorsing the workshop. In particular, the Archimedes Research Unit generously supported the participation of three of the organizers.

Last but not least, we extend our thanks to the members of the DialRes Program Committee for their contributions to the successful organization of the workshop.

The DialRes Workshop Organizers:

Antonis Anastasopoulos
Stella Markantonatou
Angela Ralli
Marcos Zampieri
Stavros Bompolas
Vivian Stamou

Organizing Committee

- Antonios Anasastasopoulos
- Stella Markantonatou
- Angela Ralli
- Marcos Zampieri
- Stavros Bompolas
- Vivian Stamou

Program Committee

- Mohamed Aghzal (George Mason University, USA)
- Syeda Sabrina Akter (George Mason University, USA)
- Esra Ayten (Istanbul Medeniyet University, Türkiye)
- Verginica Barbu-Mititelu (Romanian Academy – Research Institute for Artificial Intelligence, Romania)
- Johnatan Bonilla (Humboldt University of Berlin, Germany)
- Cristina Bosco (University of Turin, Italy)
- Stergios Chatzikyriakidis (University of Crete, Greece)
- Antonis Dimakis (National and Kapodistrian University of Athens & Archimedes Center, Greece)
- Fahim Faisal (George Mason University, USA)
- Voula Giouli (Aristotle University of Thessaloniki, Greece)
- Mirjana Ilić (University of Niš, Serbia)
- Brian Joseph (Ohio State University, USA)
- Petros Karatsareas (University of Westminster, UK)
- Panagiotis Krimpas (Democritus University of Thrace, Greece)
- J. Elizabeth Liebl (George Mason University, USA)
- Johann-Mattis List (University of Passau, Germany)
- Nikola Ljubešić (Jožef Stefan Institute, Slovenia)
- Irina Lobzhanidze (Ilia State University, Georgia)
- Chutong Meng (George Mason University, USA)
- Nikolett Mus (Hungarian Research Centre for Linguistics, Hungary)
- Petya Osenova (Sofia University "St. Kliment Ohridski", Bulgaria)

- Joshua Otten (George Mason University, USA)
- Georgios Paraskevopoulos (ILSP/Athena Research Center, Greece)
- Siyao (Logan) Peng (LMU Munich, Germany)
- Alistair Plum (University of Luxembourg, Luxembourg)
- Prokopis Prokopidis (ILSP/Athena Research Center, Greece)
- Christoph Purschke (University of Luxembourg, Luxembourg)
- Josep Quer (ICREA – Universitat Pompeu Fabra, Spain)
- Yves Scherrer (University of Oslo, Norway)
- Krešimir Sojat (University of Zagreb, Croatia)
- Aarohi Srivastava (University of Notre Dame, USA)
- Robert Sviben (Institute for the Croatian Language, Croatia)
- Chihiro Taguchi (University of Notre Dame, USA)
- Zeerak Talat (Simon Fraser University, Canada)
- Alexandros Tantos (Aristotle University of Thessaloniki, Greece)
- Belu Ticona (George Mason University, USA)
- Chara Tsoukala (ILSP/Athena Research Center, Greece)
- Socrates Vakirtzian (University of Crete, Greece)
- Alexandra Vella (University of Malta, Malta)
- Shuly Wintner (University of Haifa, Israel)
- Murong Yue (George Mason University, USA)

Table of Contents

A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry
Silvio Calderaro and Johanna Monti 1

Saar-Voice: A Multi-Speaker Saarbrücken Dialect Speech Corpus
Lena Sophie Oberkircher, Jesujoba Alabi, Dietrich Klakow and Jürgen Trouvain 12

MD_NLP: Reconstructing an Australian English Heritage Dialect Corpus from the Mitchell-Delbridge Recordings through LLM-Assisted Speaker Attribution
Steven Coats 24

Challenges in the Detection of Dialect for Historical Languages; the Case of Old Irish Text Resources
Adrian Doyle 33

Phonologically-aware Automatic Speech Recognition Evaluation of Low-Resource Languages: The Case of Basque Dialects
Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas and Inma Hernaez
. 48

Systematic Normalization of Spoken Mixed-Language, Mixed-Dialect Data
Margaret Blevins 58

Handling Cross-Dialect Syntactic Variation: a Theory-Driven Web Resource
Emanuela Li Destri, Marco Longhin, Gaia Sorge, Sofia Ferroni, Giovanni Battista Matteazzi, Andrea Artioli, Lorenzo Carletti, Federico Motta, Giuseppe Longobardi and Cristina Guardiano
. 70

Can LLM Agents Identify Spoken Dialects like a Linguist?
Tobias Bystrich, Lukas Hamm, Maria Hassan Akhter, Lea Fischbach, Lucie Flek and Akbar Karimi 83

Beyond Accuracy: Analyzing Dialect Confusion in Automatic Speech-Based Dialect Classification
Lea Fischbach, Alfred Lameli and Lucie Flek 93

FLEURS-Kobani: Extending FLEURS dataset for Northern Kurdish
Daban Q. Jaff and Mohammad Mohammadamini 104

Exploring the reusability of Northern Kurdish resources for Badini speech recognition
Mohammad Mohammadamini, Aveen Jalal Mohammed, Barzan Hussein Mohammed, Dezheen H. Abdulazeez, Imad Saeed Sadeeq, Dilgash Mohammed Salih, Amera Ismail Melhum and Abuobaida Abdullah Dheyab 110

Wancho Dialectometry: Community-created data and the Living Dictionaries project
Kellen Parker van Dam 116

Dialectometry and Evaluation of the ePark Corpus for Low-Resource Formosan Language Dialects
Henry Gagnier 124

A Dialectal Corpus for Ukrainian: Collection, Classification, and Standardization
Yuliia Frund and Sina Ahmadi 135

German Dialects Across Situations, Generations, and Regions: The REDE corpus as an Oral Resource for NLP
Hanna Fischer and Alfred Lameli 144

A Catalog of Basque Dialectal Resources: Online Collections and Standard-to-Dialectal Adaptations
Jaione Bengoetxea, Itziar Gonzalez-Dios and Rodrigo Agerri 153

WoVis: Interactive Visualization of Word Embeddings for Semantic Change in Historical and Dialectal Language Resources
Filip Miletić, Maximilian Henkel, Rene Cutura, Sophie Sadler, Quynh Quang Ngo, Michael Sedlmair and Sabine Schulte im Walde 165

Speaker Normalization via Voice Conversion Reveals a Human-Machine Dissociation in Dialect Classification
Caroline Kleen, Lea Fischbach, Akbar Karimi, Lucie Flek and Alfred Lameli 177

South Tyrolean Dialect-to-Standard Speech Translation: A Resource
Greta H. Franzini and Luca Ducceschi 188

TransVar – the Corpus for Variation and Change Study of the Historical Transcarpathian lects
Ilia Afanasev 195

The Generator-Eraser Paradox: Community Guidelines for Responsible LLM-Assisted Dialect Resource Creation
Wajdi Zaghouani 209

The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact
Thomas Schmidt, Margaret M. Blevins, Hans C. Boas and Glenn Gilbert 221

Pontic Greek in the Caucasus: an online corpus
Svetlana Berikashvili and Stavros Skopeteas 230

Meaning Over Morphology: A Multi-Metric Benchmark of LLMs for Bangla Dialect Translation
Soumik Deb Niloy, Subhey Sadi Rahman, Mahbub E Sobhani, Md. Golam Rabiul Alam, Farig Yousuf Sadeque and Md. Rezuwan Hassan 238

Sociolinguistic aspects of crowdsourcing for a vocal corpus of Alsatian
Pascale Erhart, Lucile Hamm, Sam Bigeard, Carole Werner, Malek Yaich and Slim Ouni 256

HeptaTAX: A Neuro-Symbolic Pipeline and Benchmark for Classifying 16th-Century Heptanesian Notarial Acts
Stergios Chatzikyriakidis, Eleni Karantzola and Vasiliki Makri 265

Towards Semantic Access and Interoperability in Digital Dialectal Atlases. A Case Study
Paola Marongiu and Simonetta Montemagni 274

A CLDF-Compliant Lexical Database for Modern Greek Dialects: Resource Design and Dialectometric Analysis
Stavros Bompolas, Natalia Chousou-Polydouri, Manuela Genitsaridi, Danae Karatzanou, Georgios Kostopoulos, Elena Anagnostopoulou and Dimitra Melissaropoulou 287

A Speech Resource for the Pontic Greek Dialect: Transcription Choices and Baseline ASR Evaluation
Rodanna Konstantinidou, Chara Tsoukala, Vivian Stamou, Voula Giouli and Stella Markantonatou 300

First Steps in ASR for Cypriot Greek: Challenges and Insights
Vivian Stamou, Spyros Armostis, Antigoni Klimi, Georgios Paraskevopoulos, Vassilis Katsouros and Antonios Anastasopoulos 308

Structural Divergence under Shared Language-Level Specification: Griko in Universal Dependencies
Stavros Bompolas, Emanuela Pinna, Josep Quer, Marika Lekakou and Stella Markantonatou 315

Digital Preservation of Aromanian Through Knowledge Management and Automatic Speech Recognition Evaluation
Marija Pendevska and Hristina Nastevska 327

A Novel Typology of Mutually Intelligible Words: The Case of Slavic Languages
Edward Klyshinsky and Yulia Badryzlova 337

Transfer Learning for an Endangered Slavic Variety: Dependency Parsing in Pomak Across Contact-Shaped Dialects
Sercan Karakas 345

Workshop Program

Saturday, May 16, 2026

09:00–10:30 **Session A**
Room: Room 4
Chair: Stella Markantonatou

09:10–09:20 *A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry*
Silvio Calderaro and Johanna Monti

09:20–09:30 *Saar-Voice: A Multi-Speaker Saarbrücken Dialect Speech Corpus*
Lena Sophie Oberkircher, Jesujoba Alabi, Dietrich Klakow and Jürgen Trouvain

09:30–09:40 *MD_NLP: Reconstructing an Australian English Heritage Dialect Corpus from the Mitchell-Delbridge Recordings through LLM-Assisted Speaker Attribution*
Steven Coats

09:40–09:50 *Challenges in the Detection of Dialect for Historical Languages; the Case of Old Irish Text Resources*
Adrian Doyle

10:00–10:30 ***2-minute poster presentations***

10:30–11:00 **Coffee break**

10:30–11:00 **Poster Session**
Chairs: Stavros Bompolas, Vivian Stamou

Saturday, May 16, 2026 (continued)

11:00–13:00 **Session B**
Room: Room 4
Chair: Antonios Anastasopoulos

11:00–11:10 *Phonologically-aware Automatic Speech Recognition Evaluation of Low-Resource Languages: The Case of Basque Dialects*
Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas and Inma Hernaez

11:10–11:20 *Systematic Normalization of Spoken Mixed-Language, Mixed-Dialect Data*
Margaret Blevins

11:20–11:30 *Handling Cross-Dialect Syntactic Variation: a Theory-Driven Web Resource*
Emanuela Li Destri, Marco Longhin, Gaia Sorge, Sofia Ferroni, Giovanni Battista Matteazzi, Andrea Artioli, Lorenzo Carletti, Federico Motta, Giuseppe Longobardi and Cristina Guardiano

11:30–11:40 *Can LLM Agents Identify Spoken Dialects like a Linguist?*
Tobias Bystrich, Lukas Hamm, Maria Hassan Akhter, Lea Fischbach, Lucie Flek and Akbar Karimi

11:45–12:30 *Invited talk (title to be announced)*
Prof. Barbara Plank, LMU Munich, Visiting Prof ITU Copenhagen

12:30–13:00 ***Community discussion***

10:30–11:00 ***List of Posters***

10:30–11:00 *Beyond Accuracy: Analyzing Dialect Confusion in Automatic Speech-Based Dialect Classification*
Lea Fischbach, Alfred Lameli and Lucie Flek

10:30–11:00 *FLEURS-Kobani: Extending FLEURS dataset for Northern Kurdish*
Daban Q. Jaff and Mohammad Mohammadamini

10:30–11:00 *Exploring the reusability of Northern Kurdish resources for Badini speech recognition*
Mohammad Mohammadamini, Aveen Jalal Mohammed, Barzan Hussein Mohammed, Dezheen H. Abdulazeez, Imad Saeed Sadeeq, Dilgash Mohammed Salih, Amera Ismail Melhum and Abuobaida Abdullah Dheyab

Saturday, May 16, 2026 (continued)

10:30–11:00 *Wancho Dialectometry: Community-created data and the Living Dictionaries project*
Kellen Parker van Dam

10:30–11:00 *Dialectometry and Evaluation of the ePark Corpus for Low-Resource Formosan Language Dialects*
Henry Gagnier

10:30–11:00 *A Dialectal Corpus for Ukrainian: Collection, Classification, and Standardization*
Yuliia Frund and Sina Ahmadi

10:30–11:00 *German Dialects Across Situations, Generations, and Regions: The REDE corpus as an Oral Resource for NLP*
Hanna Fischer and Alfred Lameli

10:30–11:00 *A Catalog of Basque Dialectal Resources: Online Collections and Standard-to-Dialectal Adaptations*
Jaione Bengoetxea, Itziar Gonzalez-Dios and Rodrigo Agerri

10:30–11:00 *WoVis: Interactive Visualization of Word Embeddings for Semantic Change in Historical and Dialectal Language Resources*
Filip Miletić, Maximilian Henkel, Rene Cutura, Sophie Sadler, Quynh Quang Ngo, Michael Sedlmair and Sabine Schulte im Walde

10:30–11:00 *Speaker Normalization via Voice Conversion Reveals a Human-Machine Dissociation in Dialect Classification*
Caroline Kleen, Lea Fischbach, Akbar Karimi, Lucie Flek and Alfred Lameli

10:30–11:00 *South Tyrolean Dialect-to-Standard Speech Translation: A Resource*
Greta H. Franzini and Luca Ducceschi

10:30–11:00 *TransVar – the Corpus for Variation and Change Study of the Historical Transcarpathian lects*
Ilia Afanasev

10:30–11:00 *The Generator-Eraser Paradox: Community Guidelines for Responsible LLM-Assisted Dialect Resource Creation*
Wajdi Zaghouani

10:30–11:00 *The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact*
Thomas Schmidt, Margaret M. Blevins, Hans C. Boas and Glenn Gilbert

10:30–11:00 *Pontic Greek in the Caucasus: an online corpus*
Svetlana Berikashvili and Stavros Skopeteas

Saturday, May 16, 2026 (continued)

10:30–11:00 *Meaning Over Morphology: A Multi-Metric Benchmark of LLMs for Bangla Dialect Translation*
Soumik Deb Niloy, Subhey Sadi Rahman, Mahbub E Sobhani, Md. Golam Rabiul Alam, Farig Yousuf Sadeque and Md. Rezuwan Hassan

10:30–11:00 *Sociolinguistic aspects of crowdsourcing for a vocal corpus of Alsatian*
Pascale Erhart, Lucile Hamm, Sam Bigeard, Carole Werner, Malek Yaich and Slim Ouni

10:30–11:00 *HeptaTAX: A Neuro-Symbolic Pipeline and Benchmark for Classifying 16th-Century Heptanesian Notarial Acts*
Stergios Chatzikyriakidis, Eleni Karantzola and Vasiliki Makri

10:30–11:00 *Towards Semantic Access and Interoperability in Digital Dialectal Atlases. A Case Study*
Paola Marongiu and Simonetta Montemagni

10:30–11:00 *A CLDF-Compliant Lexical Database for Modern Greek Dialects: Resource Design and Dialectometric Analysis*
Stavros Bompolas, Natalia Chousou-Polydouri, Manuela Genitsaridi, Danae Karatzanou, Georgios Kostopoulos, Elena Anagnostopoulou and Dimitra Melissaropoulou

10:30–11:00 *A Speech Resource for the Pontic Greek Dialect: Transcription Choices and Baseline ASR Evaluation*
Rodanna Konstantinidou, Chara Tsoukala, Vivian Stamou, Voula Giouli and Stella Markantonatou

10:30–11:00 *First Steps in ASR for Cypriot Greek: Challenges and Insights*
Vivian Stamou, Spyros Armostis, Antigoni Klimi, Georgios Paraskevopoulos, Vassilis Katsouros and Antonios Anastasopoulos

10:30–11:00 *Structural Divergence under Shared Language-Level Specification: Griko in Universal Dependencies*
Stavros Bompolas, Emanuela Pinna, Josep Quer, Marika Lekakou and Stella Markantonatou

10:30–11:00 *Digital Preservation of Aromanian Through Knowledge Management and Automatic Speech Recognition Evaluation*
Marija Pendevska and Hristina Nastevska

10:30–11:00 *A Novel Typology of Mutually Intelligible Words: The Case of Slavic Languages*
Edward Klyshinsky and Yulia Badryzlova

10:30–11:00 *Transfer Learning for an Endangered Slavic Variety: Dependency Parsing in Pomak Across Contact-Shaped Dialects*
Sercan Karakas

A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry

Silvio Calderaro[1,2], Johanna Monti[2]
[1]Università di Pisa, Italia
[2]Università di Napoli L'Orientale, Italia
silvio.calderaro@phd.unipi.it, jmonti@unior.it

Abstract

The growing interest of Natural Language Processing (NLP) in minority languages has not yet bridged the gap in the preservation of oral linguistic heritage. In particular, extemporaneous poetry — a performative genre based on real-time improvisation, metrical-rhetorical competence — remains a largely unexplored area of computational linguistics. This methodological gap necessitates the creation of specific resources to document and analyse the structures of improvised poetry. This is the context in which A Bolu was created, the first structured corpus of extemporaneous poetry dedicated to *cantada logudorese*, a variant of the Sardinian language. The dataset comprises 2,835 stanzas for a total of 141,321 tokens. The study presents the architecture of the corpus and applies a multidimensional analysis combining descriptive statistical indices and computational linguistics techniques to map the characteristics of the poetic text. The results indicate that the production of Sardinian extemporaneous poets is characterised by recurring patterns that support Parry and Lord's theory of formulaicity. This evidence not only provides a new key to understanding oral creativity, but also offers a significant contribution to the development of NLP tools that are more inclusive and sensitive to the specificities of less widely spoken languages.

Keywords: Corpus of minority languages, Extemporaneous poetry, Oral-formulaic patterns

1. Introduction and Background

The growing interest of the Natural Language Processing (NLP) community in low-resource languages, minority varieties, and dialects reflects a broader shift toward linguistic inclusivity. For decades, computational tools and annotated resources were concentrated on a small set of high-resource languages, leaving the vast majority of the world's linguistic heritage outside the reach of modern NLP techniques. Corpus construction efforts for minority languages must typically contend with sparse digital presence, orthographic instability, and the absence of standardized annotation frameworks, challenges well attested in comparable projects for languages such as Guarani (Chiruzzo et al., 2022), Breton (Grobol and Jouitteau, 2024), and the minority languages of Italy (Ramponi, 2024). These difficulties are further compounded when the target variety is rooted in oral and performative traditions (Zumthor, 1997; Ong, 1982) that have historically resisted stabilization into annotated digital formats.

Within this landscape, Sardinian occupies a particularly complex position. Officially recognized as a minority language under Italian national law and widely regarded as the closest living descendant of Latin (Virdis, 1988). Existing resources include SardNet (Angioni et al., 2018), a lexical resource mapping Sardinian word senses onto WordNet entries, a BERT-based Part-of-Speech tagger (Carta et al., 2025), and a nascent Automatic Speech Recognition system (Chizzoni and Vietti, 2024).

The *cantada logudoresa* is a formal poetic contest in which *cantadores* improvise verses in the Logudorese dialect under strict metrical and thematic constraints, structured around the *Otada*, an octave of eight hendecasyllabic lines following an ABABABCC or ABBABACC rhyme scheme, alongside shorter forms such as the *batorina* and the closing *dispedida*. Performers must simultaneously manage metrical and rhyming complexity, thematic coherence, and real-time argumentative exchange, all under the pressure of live performance. The preservation of this tradition has historically depended on inconsistent and scattered acts of transcription, precluding any systematic computational investigation. This paper introduces **A Bolu** (Calderaro and Monti, 2026), the first structured digital corpus of the *cantada logudoresa*, designed to fill this gap and establish a reproducible framework for its quantitative and linguistic analysis. Our contributions:

1. **Digital Preservation and Resource Creation:** We provide a high-fidelity digital repository for a vulnerable minority language tradition, preventing the loss of undocumented or fragmented transcriptions and establishing a foundation for future NLP tasks in Sardinian.

2. **Multidimensional Data Modeling:** Unlike flat-text corpora, *A Bolu* is structured to include rich metadata—such as thematic assignments, performer identifiers, and precise execution

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 1–11
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

timestamps per stanza—modeled in a hierarchical format to facilitate complex relational queries.

3. **Computational Stylistics Analysis:** We demonstrate the utility of the dataset by proposing it as a benchmark for investigating "stylistic signatures" and lexical complexity, enabling quantitative research into how real-time improvisational pressures affect the linguistic and metrical choices of the *cantadores*.

The aim of this resource is to integrate Sardinian extemporaneous poetry into the field of computational linguistics, defining a reproducible framework that promotes new avenues of study and formal analysis of structured oral traditions.

2. Methodology

This section describes the methodological approach adopted for the construction of the corpus, divided into the phases of data acquisition 2.1, archive structuring 2.2, and curation of the raw material 2.3. The central objective was to transform a heritage of oral tradition—fragmentary and discontinuous by nature—into a structured digital resource, suitable for computational processing and quantitative linguistic analysis.

2.1. Data Acquisition and Source Selection

The primary data for this study were collected programmatically from làcanas.it (Redazione Làcanas, 2014), an online newspaper dedicated to Sardinian culture, news, and identity, with an archive of improvised poems in the Sardinian language. The scraping process targeted the poems on the site to maintain the same origin as the source. The corpus was constructed according to strict criteria of linguistic and typological homogeneity to ensure the reliability of subsequent quantitative comparisons. The sources were selected based on three fundamental parameters:

1. **Transcription Quality and Metadata Richness:**produced by the official editorial staff were included in the corpus. This restriction was implemented to ensure consistency, reliability, and philological accuracy across the dataset. A primary selection criterion was the availability of essential contextual metadata, including the performance setting, the designated themes (i.e., thematic debates), and, crucially, the recorded execution time of each stanza for each poet.
2. **Generic Consistency:** To avoid stylistic bias, the dataset exclusively comprises performances belonging to the same poetic genre, specifically*(cantada logudoresa)*. This ensures that the metrical constraints and the thematic development remain constant across the entire sample.
3. **Linguistic Variety:** The selection was restricted to a single linguistic variety of the Sardinian language (*Logudorese*). This choice eliminates lexical variation due to dialectal shifts, allowing the analysis to focus strictly on the individual poets' lexical complexity and rhyming strategies.

Despite the application of these selection criteria, the corpus inevitably reflects the intrinsic challenges associated with documenting oral traditions. It should be noted that several performances within the dataset are incomplete: in some cases, individual stanzas are partially or entirely absent from the original transcriptions. These lacunae—often resulting from recording interruptions or archival deterioration—were systematically identified during the data-cleaning phase in order to prevent distortions in the statistical distribution of the linguistic metrics. Furthermore, the texts included in the corpus do not represent complete poetry contests. Rather, the materials primarily consist of the central phase of the competition, which accounts for the majority of the recorded performances. Consequently, these contest-based transcriptions frequently omit either the opening segment (*esordiu)* or the final closing exchange (*dispèdida*) of the debate.

2.2. Data Structure and Corpus Architecture

The collected performances were encoded in a structured, hierarchical format using `JSON` (JavaScript Object Notation). This data model was designed to represent the multi-layered organisation of extemporaneous poetic debates, preserving the internal segmentation of each performance while maintaining explicit links between individual stanzas and their associated metadata. Such a structure ensures both formal consistency and computational accessibility, thereby facilitating corpus querying, annotation, and quantitative analysis.

The corpus is organized into two primary levels:

- **Global Metadata:** This top-level object captures the contextual framework of the entire performance. It includes the *title* of the debate, the *source URL* for traceability, an *introductory summary* of the event, and the *central theme* (thematic dispute) assigned to the poets.
- **Transcription Units (Stanzas):** The `transcription` field contains an ordered array of stanza-level objects, each corresponding to a single metrical unit: namely an octave

(*Otada*), quatrain (*Batorina*), couplet (*Duina*), or closing/free-form stanza (*Dispèdida*). Each object encodes a discrete poetic turn and is associated with a set of structured attributes: a unique numerical `id` (ensuring sequential traceability), the `poet` identifier, the `metrics` label specifying the metrical form, the recorded `time` of execution, and an ordered list of `verse` strings representing the individual lines of the stanza.

An example of the data representation for a single stanza and its associated metadata is provided below:

```
{
  "metadata": {
    "title": "Sozu e Masala in Ballao:
    ↪  tempus e omine",
    "intro_theme": "Cando cantaian
    ↪  manu-manu Peppe Sozu e Marieddu
    ↪  Masala sa resultada de sa gara
    ↪  fut bella e assigurada in
    ↪  partenzia. [...]",
    "core_theme": "Disputa tra
    ↪  l'Eternita inarrestabile e
    ↪  l'Ingegno Creativo",
    "source":
    ↪  "https://www.lacanas.it/..."
  },
  "transcription": [
    {
      "id": 1,
      "poet": "Sozu",
      "metrics": "Otada",
      "time": "1'18",
      "verse": [
        "Ja non cherio in piata sa
        ↪  zente",
        "chi da s'atesa restet in
        ↪  fastizu.",
        "Como sighimos che babbu e che
        ↪  fizu",
        "tantu de allegrare s'ambiente",
        "ja chi su comitadu
        ↪  intelligente",
        "at apagadu su nostru disizu",
        "ch'in parte 'e unu tema mi
        ↪  collocat",
        "e a cantare su tempus mi
        ↪  tocat."
      ]
    }
  ]
}
```

This hierarchical architecture transforms the poetic performance from a static text into a multidimensional data object. By explicitly linking linguistic output to specific constraints—such as the thematic assignment, the metric of the stanzas and the execution time—the dataset allows for a granular investigation. Furthermore, this structured format ensures the corpus is fully interoperable with modern NLP pipelines, establishing a reproducible framework for the quantitative study of Sardinian oral traditions.

2.3. Data Processing and Corpus Curation

Once the raw data were extracted, a rigorous process of curation and normalization was required to transform the scraped material into a reliable research dataset. This phase involved both automated filtering and manual philological verification to address the inconsistencies inherent in a heterogeneous digital archive.

1. **Deduplication Record :** A primary challenge was the presence of duplicate performances. Many poetic debates were found to be published multiple times under slightly different titles or categorized in different sections of the source website. These redundant entries were identified and removed to ensure that the statistical analysis of lexical frequency and poet participation remained unbiased.

2. **Entity Resolution and Normalization:** To ensure that each poet's stylistic signature was correctly attributed, we performed a normalization of personal names. Variations in transcription, such as the inconsistent use of accents (e.g., *Màsala* vs. *Masala*), were reconciled to a single canonical form. This step is crucial for the subsequent calculation of individual lexical complexity and comparative stylometry.

3. **Structural Integrity and Lacunae Flagging:** Each stanza was checked automatically and manually to verify its completeness. Given the oral and often fragmented nature of the transcriptions, we adopted a symbolic tagging system within the *metrical form* metadata field to maintain the chronological sequence of the debate without compromising the linguistic statistics:

   - **Standard labels** (e.g., `otada`): Applied to complete stanzas where all verses are present.
   - **Single asterisk** (e.g., `otada*`): Indicates a partially missing stanza where only a portion of the verses was transcribed.
   - **Double asterisk** (e.g., `otada**`): Indicates a fully missing stanza, preserving its original position in the performance flow while excluding it from the textual analysis.

4. **Temporal Standardization:** The execution time for each stanza, recorded in the source as a string format (e.g., `1'00"`), was parsed and converted into a discrete numerical variable

representing total seconds. In cases where the timing was not present in the original source, or when the stanza was incomplete (as indicated by the asterisk system), a `null` value was assigned to the field.

This approach ensures that the resulting corpus is not merely a collection of texts, but a structured digital resource that preserves both the textual content and the contextual information of the poetic contests, including stanza-level metadata, performance timing, and thematic annotation.

3. Corpus statistics

The resulting dataset, **A Bolu**, to the best of our knowledge constitutes the first structured digital corpus of Sardinian extemporaneous poetry explicitly designed for computational analysis. By aggregating dispersed transcriptions and imposing a systematic structural organization, the corpus provides a solid foundation for rigorous empirical investigation not only into this minority language tradition but also into the structure and dynamics of extemporaneous poetic performance itself.

3.1. Statistical overview

The final corpus, consists of 55 digitalized poetic sessions with a specific focus on the 20th-century oral heritage. Table 1 presents a synoptic view of the corpus dimensions. The dataset comprises 2,835 stanzas distributed across 55 poems, yielding an average of 51.55 stanzas per poem. The corpus also documents its own gaps transparently: 60 stanzas lack execution time annotations, 63 are partially incomplete in their textual transcription, and 8 are entirely missing. Rather than discarding these entries, they have been retained with explicit flags to preserve the structural integrity of each poetic debate.

Feature	Value
Total number of poems (JSON files)	55
Total number of stanzas	2,835
Average stanzas per poem	51.55
Unique identified poets	8
Stanzas with missing execution time	60
Stanzas with partially incomplete	63
Stanzas with totally incomplete	8

Table 1: General statistics of the A Bolu corpus.

3.2. Poet contributions and metrical distribution

The following section examines how stanzas are distributed across poets and metrical forms within the corpus, highlighting both the dominance of certain performers and the structural variety of the poetic debate. Table 2 provides a comprehensive breakdown of stanza counts per poet, disaggregated by metrical type and transcription completeness. Two figures dominate the dataset: Sozu and Masala each appear in 32 poems and contribute 811 and 753 stanzas respectively, together accounting for over 55% of the total corpus. This concentration reflects their historical prominence within the tradition and ensures substantial material for in-depth stylistic analysis. Piras (493 stanzas, 20 poems) Mura (339 stanzas, 15 poems) and Piredda (246 stanzas, 10 poems) constitute a second tier, while the remaining three poets, Sale, Seu, and Budrone, contribute smaller but nonetheless valuable samples, particularly for contrastive purposes. Regarding metrical variety, the *Otada* is overwhelmingly the foundational unit of the corpus, comprising 2,592 of the 2,835 stanzas (93.9%). The remaining forms,*Duina* (95 stanzas, 3.4%), *Batorina* (70 stanzas, 2.5%), and *Despedida* (7 stanzas, 0.2%), appear with considerably lower frequency, consistent with their specialized roles within the structure of the poetic debate. The distribution of these secondary forms is not uniform across the poets. *Masala* employs the widest metrical repertoire, featuring 49 *Duinas*, 35 *Batorinas*, and 3 *Despedidas*. This is followed by *Sozu* (20 *Duinas*, 16 *Batorinas*, and 2 *Despedidas*), *Mura* (16 *Duinas*, 7 *Batorinas*, and 1 *Despedida*), and *Seu* (10 *Duinas*, 12 *Batorinas*, and 1 *Despedida*). In contrast, *Piras*, *Piredda*, *Sale*, and *Budrone* perform exclusively in the *Otada* form. This asymmetry is likely attributable to the uneven distribution of available data across poets in the original source. The specifically tagged entries *Otada** (63 stanzas) and *Otada*** (8 stanzas) denote partially and totally incomplete transcriptions, respectively. These lacunae are distributed unevenly across poets: Sozu accounts for the largest share of partial gaps (21 *Otada**), a figure that is proportional to his extensive presence in the corpus and likely reflects the variable quality of the source recordings. By retaining these entries with explicit markers rather than omitting them, the dataset preserves the sequential structure and turn-taking logic of each session, which is essential for any analysis of debate dynamics and interactional patterns between competing poets.

3.3. Lexical analysis

In the context of oral improvised poetry, the lexical dimension acquires particular significance: poets must generate verse in real time under strict metrical and rhyming constraints, drawing on a mental lexicon that is simultaneously broad enough to avoid repetition and sufficiently controlled to satisfy

Poet	Tot.	Otada	Otada*	Otada**	Duina	Bator.	Disped.	Poems
Sozu	811	751	21	1	20	16	2	32
Masala	753	656	6	4	49	35	3	32
Piras	493	472	18	3	0	0	0	20
Mura	339	305	10	0	16	7	1	15
Piredda	246	242	4	0	0	0	0	10
Sale	88	85	3	0	0	0	0	4
Seu	86	62	1	0	10	12	1	4
Budrone	19	19	0	0	0	0	0	1
Total	**2,835**	**2,592**	**63**	**8**	**95**	**70**	**7**	**-**

Table 2: Comprehensive breakdown of stanzas per poet, including metrical forms and transcription lacunae (*Otada** = partially incomplete; *Otada*** = totally incomplete).

the formal requirements of the tradition. Measuring lexical richness in such a setting can shed light on the cognitive and linguistic resources that differentiate one performer from another. To ensure a methodologically robust evaluation and to mitigate the well-documented sensitivity of the standard Type-Token Ratio (TTR) to corpus length (Baayen, 2001), we employ two length-independent indices: the Moving Average Type-Token Ratio (MATTR) (Covington and McFall, 2010) and the Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010). For the MATTR calculation, a sliding context window of 50 words was selected, as it approximates the average length of a single *ottava*, the fundamental metrical unit of the performance, thus capturing lexical density at the level of individual poetic turns.

Table 3 reports token counts, type counts, hapax legomena, and the calculated lexical diversity indices for each poet. The full corpus comprises 141,321 tokens and 12,973 types, yielding an overall TTR of 9.18%. This low figure is consistent with the lexical saturation expected in a genre-constrained oral corpus, where recurrent grammatical forms and thematic vocabulary inevitably accumulate across large text volumes. Individual sub-corpora range from 968 tokens (Budrone) to 41,164 tokens (Sozu), a difference of more than fortyfold, which makes direct TTR comparisons across poets methodologically unreliable.

As expected, raw TTR values vary considerably, from 14.47% (Sozu) to 44.01% (Budrone), a divergence that reflects corpus size rather than genuine differences in lexical competence. When examining the MATTR (window = 50), however, scores are remarkably stable across all eight performers, with values ranging from 79.98 (Budrone) to 83.26 (Mura) and a mean of $\mu = 81.59$. This convergence suggests that local lexical density – measured within windows that approximate the length of a single *ottava* – constitutes a structural property of the *Cantada a bolu* tradition, largely independent of the total volume of text produced by each poet.

The MTLD scores provide a complementary perspective on sustained lexical diversity. The global MTLD computed on the entire concatenated corpus is 95.87, with individual values ranging from 81.20 (Budrone) to 110.77 (Mura). These figures indicate a consistent capacity for lexical variation throughout extended stretches of discourse. Mura's score of 110.77 stands out as the highest in the sample, suggesting an exceptional ability to vary vocabulary over longer sequences, even when compared to poets with larger sub-corpora such as Sozu or Masala.

The hapax legomena counts offer a further dimension of analysis. Across the full corpus, 6,789 word forms occur exactly once, accounting for approximately 52.33% of the 12,973 distinct types attested. This substantial proportion of singletons suggests that the poets do not rely exclusively on a fixed repertoire of formulaic expressions but instead introduce novel vocabulary throughout their performances. Taken together, the MATTR and MTLD results indicate that the *A Bolu* corpus exhibits a high and consistent level of lexical richness across sub-corpora of considerably different sizes, lending support to the view that local lexical density is a stable feature of this oral poetic tradition.

Author	Tokens	Types	Hapax	TTR%	MATTR	MTLD
Sozu	41,164	5,956	3,295	14.47	81.89	97.13
Masala	36,305	5,779	3,411	15.92	81.80	98.82
Piras	25,573	4,435	2,712	17.34	80.56	85.65
Mura	16,507	3,302	2,053	20.00	83.26	110.77
Piredda	12,440	2,701	1,674	21.71	80.79	94.85
Sale	4,576	1,344	915	29.37	82.60	100.10
Seu	3,788	1,226	825	32.37	82.40	104.93
Budrone	968	426	301	44.01	79.98	81.20
Total	**141,321**	**12,973**	**6,789**	**9.18**	**81.59**	**95.87**

Table 3: Lexical statistics per author and corpus totals: tokens, types, hapax legomena, TTR, MATTR, and MTLD.

3.4. Temporal Dynamics and Metric Complexity

The inclusion of execution time as a granular attribute introduces an important dimension for analy-

sis in this specific domain. Table 4 summarizes the average duration (in seconds) for the three primary metrical forms identified: the *Otada*, the *Duina*, and the *Batorina*.

The data reveal a clear correlation between metrical length and temporal duration. The *Otada*, being the foundational and most complex unit of the *cantada*, requires a corpus-wide average of 58.18 seconds per unit. In contrast, the *Duina* and the *Batorina* function as more rapid improvisational units,due to the brevity of their short nature, with average execution times of 10.02 and 24.93 seconds, respectively.

An analysis of individual poet performances highlights significant variations in improvisational pace:

- **Masala** emerges as the most rapid improviser across all categories, with an average *Otada* time of 45.57 seconds, significantly below the general mean of 58.18 seconds.
- **Sozu**, despite being one of the most prolific contributors, maintains a more measured pace with an average of 68.33 seconds per *Otada*.
- **Sale** recorded the highest average time for a complete octave (83.25 seconds), suggesting a different stylistic approach to the rhythmic constraints of the performance.

It should be noted that these poetic performances are mainly sung and often accompanied by musical interludes. For this reason, it would be misleading to correlate temporal variations with lexical data, as this would not provide truly meaningful information about the creative abilities of the poets. Some poets, in fact, prefer to dwell on the vowels /a/, /e/ and /ɛ//.

For this reason, temporal characteristics should be considered as stylistic traits specific to each poet, rather than as descriptors of actual improvisational abilities.

Author	Otada (s)	Duina (s)	Batorina (s)
Sozu	68.33	11.55	26.69
Masala	45.57	8.51	20.80
Piras	53.83	–	–
Mura	65.41	12.00	31.71
Seu	66.95	10.90	30.67
Sale	83.25	–	–
Piredda	49.20	–	–
Budrone	56.89	–	–
Mean (All)	**58.18**	**10.02**	**24.93**

Table 4: Average execution time (seconds) for the primary metrical forms per author.

3.5. Co-occurrence and Formularity

To investigate recurrent multiword patterns in the corpus, we conducted an exhaustive n-gram analysis with n ranging from 3 to 8. The analysis is exploratory in nature: its aim is not to establish statistically validated categories of formulaic reuse, but to identify recurring multiword patterns and examine their functional distribution across poets and performance events. The analysis was carried out on the full transcribed corpus of 135,747 tokens and 12,808 unique word forms, comprising performances exclusively in the *ottava* meter (*otada*). Token extraction followed a conservative normalization pipeline: lowercase conversion, removal of punctuation while preserving apostrophized clitics, and whitespace normalization, with original surface forms retained separately for display purposes. Shorter sequences ($n \leq 4$) are more likely to include high-frequency grammatical collocations and are therefore less diagnostically reliable in isolation; sequences of five tokens or more, especially when recurring across distinct performance events or poets, provide stronger evidence of formulaic reuse, with higher-order sequences ($n \geq 6$) offering the most robust signal, being least likely to occur by chance.

For each candidate sequence, two association measures were computed. Pointwise Mutual Information (Church and Hanks, 1990) is defined as:

$$\text{PMI} = \log_2 \frac{P(w_1, \ldots, w_n)}{\prod_{i=1}^{n} P(w_i)}$$

where $P(w_1, \ldots, w_n)$ is the observed joint probability of the sequence and $\prod P(w_i)$ is the expected probability under lexical independence. PMI is a reliable indicator of lexical *distinctiveness* but unstable at low observed frequencies, as sequences containing rare items yield inflated values. It was therefore supplemented with the Log-Likelihood Ratio (Dunning, 1993), G^2:

$$G^2 = 2 \sum_i O_i \ln \frac{O_i}{E_i}$$

where O_i and E_i are observed and expected cell counts contrasting occurrences of the sequence against all other corpus positions. G^2 is more stable at low frequencies and penalizes sequences whose association score is driven by constituent rarity rather than genuine co-occurrence preference. Since all n-gram candidates were evaluated within a single exploratory pass over the corpus, G^2 values are not interpreted as hypothesis tests: multiple comparison corrections would be required for inferential validity, and given the exploratory nature of the analysis we adopt G^2 strictly as a *relative ranking measure* of lexical cohesion.

The verses considered are reported in Table 5, grouped into four functional categories assigned on the basis of qualitative distributional and pragmatic criteria, reflecting the different roles a verse can play within the overall structure of a performance.

Verse / Template	Poet(s)	n	Obs.	PMI	LLR (G^2)	Type
FC 1 — Deictic formulas						
de fronte a unu pópulu [ADJ] (transl. *in front of a [ADJ] people*)	Masala	5	8	27.29	286.68	Personal deixis formula
de su palcu in susu (transl. *from the stage up above*)	Mura / Masala	5	4	22.95	119.27	Shared deixis formula
FC 2 — Personal turn-management formulas						
[ADV/PRON] una cosa narrer ti cheria (transl. *[ADV/PRON] one thing I would like to tell you*)	Sozu	5	5	29.44	194.09	Personal formula (var. onset)
su chi ses nelzende pone cura (transl. *pay attention to what you are (weaving) saying*)	Piredda	6	3	42.72	171.66	Personal formula
FC 3 — Dialogic mirroring						
a morrer de piumu est un'onore (transl. *to die by lead (a bullet) is an honour*)	Piras → Sozu	8	2	46.62	125.25	Dialogic mirroring
chi si l'at mandigada sa rustia (transl. *that the blight has devoured it*)	Masala → Mura	7	2	45.50	122.15	Dialogic mirroring
FC 4 — Formulaic variants & cross-poet templates						
e cando mai chi non b'as cumpresu (transl. *and when have you ever not understood it*)	Piredda	8	2	45.80	122.98	Cross-poet template
e cando mai chi non l'as cumpresa (transl. *and when have you ever not understood it*)	Sozu	8	2	45.57	122.34	Cross-poet template
[V] est chi non l'as [V-PART] (transl. *[it] is that you have not [V-PART] it*)	Mura / Sale	5	3	16.71	63.48	Cross-poet template

Table 5: Recurring n-grams ($n = 5$–8) grouped by functional category (FC), ordered by LLR within each group. *Obs.* = observed frequency; PMI = Pointwise Mutual Information ($\log_2$); LLR = Log-Likelihood Ratio (G^2), used as a relative ranking measure; no significance thresholds are applied. Square brackets denote variable slots.

The **first group** illustrates deictic formulas with productive open slots. The 5-gram *de fronte a unu pópulu [ADJ]* ($PMI = 27.29$, $G^2 = 286.68$, obs. 8) is attested across eight distinct performances of Masala with five adjectival completions (*signore*, *devotu*, *gentile*, *'e zente*, *festosu*), and is classified as an idiolectal formula. The 5-gram *de su palcu in susu* ($PMI = 22.95$, $G^2 = 119.27$, obs. 4) is distributed across Mura and Masala in four distinct performances, with a variable left context (*da chi t'agatas*, *las improviso*, *chi as cantadu*, *da chi m'agato*) and a frozen deictic core, indicating membership in the shared traditional repertoire. These two sequences illustrate a gradient from idiolectal formulas with a productive terminal slot to cross-poet templates with a productive left context.

The two instances grouped under the **second group** represent a functionally coherent class of *personal formulas* used to manage the poet's own turn at the opening or early internal position of a strophe. Both n-grams are strictly monoauctorial — all attested occurrences are attributed to a single poet — and both carry a consistent pragmatic function across distinct performance events and distinct opponents. The first verse is the 5-gram *[ADV/PRON] una cosa narrer ti cheria* ($PMI = 29.44$, $G^2 = 194.09$, obs. 5), attested across five distinct performances of Sozu from 1958 to the early 1980s with three onset realizations (*ma*, *deo*, *solu*). The five-word core is lexically frozen while the initial position accommodates metrically interchangeable elements. The formula marks a discourse-level shift to direct personal address and is exclusive to Sozu across the corpus, constituting evidence of idiolectal formulaic stability over a long temporal span. The second verse, *su chi ses nelzende pone cura* (transl. *pay attention to what you are weaving*; $n = 6$, obs. $= 3$, $PMI = 42.72$, $G^2 = 171.66$), is exclusively associated with Piredda and distributed across three distinct performance events (1976 and twice in 1980) involving three distinct opponents, suggesting idiolectal stability over at least a four-year span.

The third group collects instances of dialogic mirroring, attested here by two examples. The 8-gram *a morrer de piumu est un'onore* ($PMI = 46.62$, $G^2 = 125.25$, obs. 2) occurs in two consecutive strophes of the same performance (see Table 6 in the Appendix): Piras deploys it as the closing verse of strophe #62 and Sozu opens his immediate response (strophe #63) with that same line, recontextualizing the formula to dismantle the argument

it had supported. Where Piras invokes dying by bullet as an honor applicable to the bandit, Sozu accepts the formula but redirects it to argue that such honor is not indiscriminate. The second case, the 7-gram *chi si l'at mandigada sa rustia* (strophes #8–#9, Masala $\rightarrow$ Mura, $PMI = 45.50$, $G^2 = 122.15$, obs. 2 see; Table 7 in the Appendix)), follows the same structural logic. Masala introduces the verse as a closing statement on the poor harvest, and Mura opens the following strophe by echoing it verbatim before expanding the causal frame, adding the *peronòspera* (downy mildew) as a further agent of destruction. Both cases instantiate verbatim n-gram repetition across a turn boundary, yet in functionally distinct ways: argumentative inversion, in which the repeated formula is turned against its original claim, and elaborative extension, in which it is accepted and expanded with additional causal content.

The **fourth functional group** collects sequences that instantiate the same pragmatic act — asserting the opponent's failure to understand — at different levels of lexical fixity. The 8-gram *e cando mai chi non b'as cumpresu*(Piredda, $PMI = 45.80$, $G^2 = 122.98$, obs. 2) and *e cando mai chi non l'as cumpresa* (Sozu, $PMI = 45.57$, $G^2 = 122.34$, obs. 2) are near-minimal pairs sharing the invariant core *cando mai chi non [PRON] as cumpress-*, differing only in clitic form and participial gender inflection. Their attribution to two distinct poets indicates a cross-poet template. The 5-gram *est chi non l'as [V-PART]* ($PMI = 16.71$, $G^2 = 63.48$, obs. 3), distributed across Mura and Sale with completions *cumpresa* and *iscritu*, represents a more abstract level: a productive syntactic frame with an open participial slot.

4. Discussion

This study aimed to address two interconnected objectives: the construction of a structured digital corpus of Sardinian extemporaneous poetry suitable for computational analysis, and the empirical investigation of formulaic behavior within this tradition.

On the first front, **A Bolu** represents, to our knowledge, the first resource of its kind for the *cantada logudoresa*. Its hierarchical architecture demonstrates that the fragmentary and discontinuous nature of oral tradition heritage can be transformed into a structured and computationally accessible dataset without sacrificing contextual richness.

Lexical and temporal analyses provide descriptive baselines that characterize the individual stylistic profiles of the eight poets. The shared lexical core that emerges from the aggregate count of types, combined with individual variation in hapax legomena and performance tenses, suggests that the poets operate within a common expressive framework while maintaining recognizable personal signatures. These findings provide the necessary empirical grounding for the more theoretically significant question of formularity: the observed n-gram recurrences are not an artifact of lexical poverty but reflect genuine formulaic preference within a rich and varied individual repertoire.

On this front, the n-gram analysis yields the most significant results. The recurrence of higher-order n grams across distinct competitions and the positional coherence of reused lines suggest the existence of a formulaic layer in the compositional process of the *cantada logudoresa*, consistent with the **Oral-Formulaic Theory** developed by Parry and Lord (Parry and Parry, 1987; Lord, 1995). Originally conceived to account for the compositional mechanisms of Homeric epic poetry, the theory posits that oral poets rely on a shared formulaic repertoire to navigate complex metrical constraints under real-time cognitive pressure. In the context of the *cantada logudoresa*, such formulas serve a dual function: they operate as mnemonic anchors and as structural building blocks, enabling the improvising poet to construct metrically and thematically coherent octaves while responding spontaneously to a rival.

The evidence further points to a layered architecture of formulaic competence, operating simultaneously at the collective and individual level. At the collective level, cross-poet templates indicate membership in a shared traditional repertoire. At the individual level, strictly monoauctorial formulas — stable across different opponents and performance events spanning years or decades — suggest that formulaic competence also has a deeply personal dimension, individually cultivated over long stretches of a poet's career, in line with the idiolectal dimension of oral-formulaic composition discussed in Foley (Foley, 1988).

An interesting result is the interactional dimension of this formularity. The reuse of a rival's line as the opening of one's own strophe suggests that formulas function not only as individual mnemonic devices, but also as competitive rhetorical strategies — a dimension that the classical Oral-Formulaic Theory, developed primarily in the context of monologic epic composition, does not explicitly account for. Two distinct modes are attested: argumentative inversion, in which the repeated formula is turned against its original claim, and elaborative extension, in which it is accepted and expanded with additional content. This finding may point to a distinctive feature of agonistic improvisational traditions more broadly.

Overall, the data are consistent with the plausibility of the oral-formulaic hypothesis applied to this tradition. However, given the exploratory nature of

the analysis, these results should be interpreted as stimuli for further investigation rather than as a formal test of the theory. Larger datasets and methods capable of identifying near-variant formulas beyond exact string matching will be needed to move from suggestive evidence to robust empirical validation.

5. Conclusion and Future Works

This study has introduced **A Bolu**, the first structured digital corpus of Sardinian extemporaneous poetry designed for computational analysis, and has presented a preliminary investigation of lexical, temporal, and formulaic dimensions of the *cantada logudoresa*. The results provide preliminary empirical support for the oral-formulaic hypothesis within this tradition, and reveal an interactional dimension of formulaic reuse that warrants further theoretical attention.

Several directions for future work follow naturally from the limitations of the present study. The most pressing priority is the expansion of the corpus itself: incorporating new poetic sessions and a broader range of poets would substantially increase the statistical power of the analysis and improve the generalizability of the findings beyond the eight performers currently represented. A larger and more diverse dataset would also allow for more reliable cross-poet comparisons, which are currently constrained by the pronounced imbalance in subcorpus sizes.

A particularly promising avenue is the integration of the audio dimension. The current corpus is exclusively text-based, which means that relevant prosodic and musical features of the performance — including melodic contour, vowel prolongation, and rhythmic patterning — remain outside the scope of the analysis. The inclusion of aligned audio recordings would not only allow for a more accurate interpretation of execution time data, but would also open the way for acoustic and phonetic analyses capable of capturing aspects of improvisational competence that are invisible at the textual level.

Finally, the development of more sophisticated metrics for evaluating stanza complexity represents a critical methodological challenge for future work. The descriptive indices used here — hapax counts, raw n-gram frequencies, TTR, MATTR and MTLD — provide a useful baseline but are insufficient to capture the full richness of the stylistic, linguistic, and cognitive phenomena at play in real-time oral composition. Length-normalized diversity metrics, network-based models of inter-poet formulaic exchange, and methods for detecting near-variant formulas beyond exact string matching would all contribute to a more nuanced understanding of how individual poets navigate the tension between metrical constraint, thematic coherence, and improvisational speed. Ultimately, such developments would position the **A Bolu** corpus as a reference resource not only for the study of Sardinian oral traditions, but for the broader computational investigation of extemporaneous poetic performance as a window into the stylistic, linguistic, and cognitive dimensions of human creativity under constraint.

Acknowledgements

The authors wish to thank the editorial staff of *Làcanas* and the publisher *Domus de Jana* for their kind availability and for granting permission to use their data in the construction of this corpus.

6. Bibliographical References

Manuela Angioni, Franco Tuveri, Maurizio Virdis, Laura Lucia Lai, and Micol Elisa Maltesi. 2018. SardaNet: A linguistic resource for Sardinian language. In *Proceedings of the 9th Global WordNet Conference*, pages 412–419, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

R Harald Baayen. 2001. *Word frequency distributions*, volume 18. Springer Science & Business Media.

Salvatore Mario Carta, Stefano Chessa, Giulia Contu, Andrea Corriga, Andrea Deidda, Gianni Fenu, Luca Frigau, Alessandro Giuliani, Luca Grassi, Marco Manolo Manca, Mirko Marras, Francesco Mola, Bastianino Mossa, Paola Mura, Marco Ortu, Leonardo Piano, Simone Pisano, Antonella Pisu, Alessandro Sebastian Podda, Livio Pompianu, Sara Seu, and Simona Giuseppina Tiddia. 2025. A BERT-based approach for Part-of-Speech tagging in the Sardinian language. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*. CEUR-WS.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel Guarani-Spanish corpus for MT benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107, Marseille, France. European Language Resources Association.

Ilaria Chizzoni and Alessandro Vietti. 2024. Towards an ASR system for documenting endangered languages: A preliminary study on Sardinian. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, volume 3878 of *CEUR Workshop Proceedings*, Pisa, Italy. CEUR-WS.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

John Miles Foley. 1988. *The Theory of Oral Composition: History and Methodology*. Indiana University Press, Bloomington, IN.

Loïc Grobol and Mélanie Jouitteau. 2024. ARBRES Kenstur: A Breton-French parallel corpus rooted in field linguistics. In *Proceedings of the Fourteenth Language Resources and Evaluation Conference*, Torino, Italy. European Language Resources Association.

Albert Bates Lord. 1995. *The singer resumes the tale*. Cornell University Press.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Walter J. Ong. 1982. *Orality and Literacy: The Technologizing of the Word*. Routledge, London.

Milman Parry and Adam Parry. 1987. *The making of Homeric verse: The collected papers of Milman Parry*. Oxford University Press.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38. First appeared as arXiv preprint arXiv:2209.09757, September 2022.

Redazione Làcanas. 2014. Poetas a bolu. Rubrica online, *Làcanas – Rivista bilingue delle identità*. Accessed: 2025. Available at: `https://www.lacanas.it/rubrica/poetas-a-bolu/`.

Maurizio Virdis. 1988. Sardisch: Areallinguistik. In Günter Holtus, Michael Metzeltin, and Christian Schmitt, editors, *Lexikon der Romanistischen Linguistik (LRL)*, volume 4, pages 897–913. Max Niemeyer Verlag, Tübingen.

Paul Zumthor. 1997. The construction of orality. *Poetry in Speech: Orality and Homeric Discourse*, page 7.

7. Language Resource References

Language Resources

Silvio Calderaro and Johanna Monti. 2026. *A Bolu: a Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry*. PID https://doi.org/10.5281/zenodo.19264263.

Appendix A. Dialogic Mirroring: Full Stanzas and English Translations

Original Sardinian Text	
Stanza #62 (Piras)	**Stanza #63 (Sozu)**
Ma cussos una fama tenen totu e sempre de sa zente sun a galla. A tie fatu t'an sa facia gialla proite su valore nde as connotu. Nara proite su Milite Ignotu no fut bandidu e moltu l'an a balla. Est moltu che bandidu e gherradore: **a morrer de piumu est un'onore.**	**A morrer de piumu est un'onore** ma cuss'onore non deghet a totu prite disparidade no as connotu o lis pones su propiu valore chi a fiancu 'e Marras e Pintore mi cheres ponner su Milite Ignotu ch'in su Vitorianu che at sa losa: creo no siat sa matessi cosa.
English Translation	
But they all have a reputation *and are always held high by the people.* *To you, they made your face turn pale* *because you have known their value.* *Tell me why the Unknown Soldier* *was not a bandit, yet he was killed by a bullet.* *He died like a bandit and a warrior:* ***to die by lead is an honor.***	***To die by lead is an honor*** *but that honor does not suit everyone* *because you have seen no disparity* *or you give them the same value* *when, alongside Marras and Pintore,* *you want to place the Unknown Soldier* *who has his tomb in the Vittoriano:* *I believe it is not the same thing.*

Table 6: Comparison of stanzas for the *piumu* occurrence.

Original Sardinian Text	
Stanza #8 (Masala)	**Stanza #9 (Mura)**
S'annu passadu s'annada fit mala, canta sicagna e canta caristia! In tota sa Sardigna de ua ebbia si nd'at salvadu solu calchi iscala: fit cosa in generale e in-d-ogni ala **chi si l'at mandigada sa rustia.** Ma bell'e gai s'oju nd'at annotu chi su tantu 'e su 'inu ch'est etotu.	**Chi si l'at mandigada sa rustia** tue afirmadu as che poesianu. No fit s'annu passadu cuss'ebbia, bistadu est su tribagliu totu invanu e ocannu ca est fritu su 'eranu sa 'ide l'at distruta sa 'iddia: s'annu passadu sende a bide pròspera mandigada si l'at sa peronòspera.
English Translation	
Last year the season was bad, *so much drought and so much famine!* *In all of Sardinia, only a few clusters* *of grapes were saved here and there:* *it was a general thing in every part* ***that the frost has eaten it all up.*** *But even so, the eye notices* *that the amount of wine is still the same.*	***That the frost has eaten it all up*** *you have stated like a poet.* *It wasn't just like that last year,* *all the labor was in vain* *and this year, because the spring is cold,* *the frost has destroyed the vine:* *last year, while the vine was prosperous,* *the downy mildew ate it up.*

Table 7: Comparison of stanzas for the *rustia* occurrence.

Saar-Voice: A Multi-Speaker Saarbrücken Dialect Speech Corpus

Lena S. Oberkircher, Jesujoba O. Alabi, Dietrich Klakow, Jürgen Trouvain
Language Science and Technology (LST), Saarland University,
Saarbrücken, Germany
{s8leober@stud, jalabi@lsv, dietrich.klakow@lsv, trouvain@lst}.uni-saarland.de

Abstract

Natural language processing (NLP) and speech technologies have made significant progress in recent years; however, they remain largely focused on standardized language varieties. Dialects, despite their cultural significance and widespread use, are underrepresented in linguistic resources and computational models, resulting in performance disparities. To address this gap, we introduce **Saar-Voice**, a six-hour speech corpus for the Saarbrücken dialect of German. The dataset was created by first collecting text through digitized books and locally sourced materials. A subset of this text was recorded by nine speakers, and we conducted analyses on both the textual and speech components to assess the dataset's characteristics and quality. We discuss methodological challenges related to orthographic and speaker variation, and explore grapheme-to-phoneme (G2P) conversion. The resulting corpus provides aligned textual and audio representations. This serves as a foundation for future research on dialect-aware text-to-speech (TTS), particularly in low-resource scenarios, including zero-shot and few-shot model adaptation.

Keywords: TTS, German, Saarbrücken, Saarland, Low-resourced Variety

1. Introduction and Background

The field of natural language processing (NLP) has seen substantial advances in recent years, driven by large-scale models and increased computational resources. Nevertheless, research and tools remain heavily focused on standardized language varieties, leaving dialectal varieties underrepresented despite their widespread use and importance to speakers' cultural identity (Blasi et al., 2022). In Germany, for example, more than 40% of the population regularly speaks regional dialects, yet these varieties are often subject to social stereotyping (Adler and Hansen, 2022). This marginalization is reflected in the technological landscape: large language models (LLMs) and speech processing systems frequently struggle with dialectal variation, as they are predominantly trained on standardized language data (Krückl et al., 2025). Consequently, recent years have seen growing efforts to develop language resources for dialectal varieties (Blaschke et al., 2024; Faisal et al., 2024; Blaschke et al., 2025) and to systematically evaluate the performance of existing NLP systems on such data (Bui et al., 2025; Muñoz-Ortiz et al., 2025).

In this paper, we contribute to this line of work by focusing on the Saarbrücken dialect of German spoken in the state of Saarland through the creation of **Saar-Voice**, a six-hour multi-speaker speech corpus.[1] In the following subsections, we provide some basic background about the Saarbrücken dialect, including the main linguistic characteristics of the dialect.

[1]The dataset is available on https://huggingface.co/datasets/UdS-LSV/Saar-Voice.

1.1. Dialect areas

The Saarland is one of Germany's 16 federal states (Bundesländer), located in the southwest of the country and bordering France and Luxembourg. The Saarland has around 1 million inhabitants. The state is, much like the rest of Germany, a linguistically diverse region, with many variants of the standard language, from regiolects to dialects.

A typical problem in linguistics is defining which exact dialect is meant with a given name. In the region of the Saarland, the broad range of regiolects and dialects are loosely referred to as "Saarländisch", but this is not a precise term. On a large scale, "Saarländisch" consists of two different dialect areas: Rhine Franconian and Moselle Franconian (see Figure 1). Both dialect areas are rather diverse on all linguistic levels, mainly phonology, morpho-syntax and lexicon. Looking more closely, there is large variance even within the two regions.

We decided in this paper to consider the Rhine Franconian dialect spoken in and around the capital of and largest city in the Saarland, Saarbrücken, for which a grammar (Steitz, 1981) and a suggested dictionary (Braun and Mangold, 1984) exist, both authored by linguists, as well as a community of active authors (with no background in linguistics).

1.2. Spelling

One big advantage of a standard language is the existence of a standard spelling. For (German) dialects standardized orthographies are missing, which leads to high variance in spelling between

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 12–23
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026



Figure 1: The isogloss *dat-das-line* distinguishing areas in which Rhine Franconian dialects ("das") and Moselle Franconian dialects ("dat") are spoken (Drenda, 2008) in Saarland (smaller area in the south-west marked with thin border lines) and Rhineland-Palatinate.

and even within speakers. Often, speakers use spontaneous spelling, orthographically representing the phonology each word as they best see fit. For instance, the exact spelling of "es gebbd" (e. 'there is', 'there are') can vary a lot: "es gebbd, gebd, gäbbd, gäbd, gebbt, gäbt"; or "Zeidung" (e. "newspaper') as "Zeidung, Tseidung, Dsaidung".

Generally, speakers use the letters available in the Standard German alphabet ("a-z, ä, ö, ü" and "ß"). Additionally, depending on the speaker, certain diacritics that are not part of the standard language may be used. For example, "ò" may be used to represent the phoneme /ɔ/ (such as in "dò", "dòò", "do" or "doo", all representing /dɔː/, meaning "there"), and in some cases *á* to represent the phoneme /ɐ/ as in "vákaafe" (/fɐkaːfə/, e. 'sell').

1.3. Differences between dialect and standard language

Were dialects simply accents of the standard language, the approach to creating any speech system would be to simply 'translate' the standard pronunciation to the regional pronunciation (ideally with the intonation of the accent). However, dialects are not just the standard language with regional accents. In fact, they often differ to written standard language on various levels:

- **Orthography and phonology:** Dialectal writing reflects phonological distinctions that are not represented in Standard German orthography. For example, vowel quality and length may be encoded differently, and consonant correspondences do not always align with standard spelling conventions.

- **Morpho-syntax:** Grammatical markers such as case or gender markings may differ between the dialect and standard language, among other morpho-syntactic differences.

  – Standard German: *ein* (indefinite article, masc./neut.) + *schön* (nice) + *Überraschung* (surprise; fem.) → *eine schöne Überraschung*

  – Dialect: *e* (indefinite article, generalized form) + *scheen* (nice) + *lwwerraschung* (surprise) → *e scheeni lwwerraschung*

  Here, both the article system and adjectival inflection differ from Standard German.

- **Lexical variation:** Dialects employ vocabulary items that are absent from Standard German or derived from different etymological sources, e.g., *Gellerieb* ('carrot', from "gelbe Rübe", in Standard German "Karotte" or "Möhre"), *Grumbier* ('potato', from "Grundbirne", in Standard German "Kartoffel"), or *Troddwaa* ('sidewalk', from French "trottoir", in Standard German "Gehsteig" or "Bürgersteig").

An important feature of dialect texts is that they are rarely found in the everyday 'linguistic landscape'. This has not only consequences for feelings of insecurity of spelling among speakers, including inconsistencies of spelling of identical words in the same text, but also for potential readers. Most *speakers* of dialects are not very good *readers* of dialect text, though they are very fluent and competent readers of standard language (but maybe speakers with a regional accent).

1.4. Text sources

There are different sources of dialect-writing, for example:

- web-based: chats, blogs
- books written by (non-professional) dialect authors
- research pieces, e.g. dictionaries, monographs and translations of local dialects
- texts from authors organized in dialect clubs
- public media, e.g. news in local radio stations, glosses in local newspaper

Source	Domain	# Sent	# Words	# U. Words
Printed Books				
An da Saar gefonn	Poetry	698	6,228	1,978
Geschaffd - Geläädbd	Poetry; Prose	394	3,302	1,214
Saa, was de willschd	Prose; Poetry	2,216	19,711	3,147
Was wääs dann isch	Poetry; Prose	2,525	18,445	3,490
Locally Sourced				
(Locally Sourced Texts)	Prose; Folktale; Poetry	2,838	26,789	5,221
Localized Translation				
MASSIVE German	Localized Data	101	805	394
Total		8,772	75,280	11,303

Table 1: Textual corpus statistics of the collected texts for the Saarbrücken dialect.

Depending on selected types of text, the styles can hugely differ. For instance, many dialect writers produce poems, in which the stanzas often rhyme at the end. In contrast, in chats reduced forms of written conversations can be observed which are again quite different to some dialect prose. Thus, the selected text source used for later steps can have a strong bias of style. This is especially important when creating a speech corpus, which may in the end contain biases in pronunciation and intonation.

2. Related Work

German Speech Resources Several resources exist for German spoken language research, but most are limited in dialectal coverage. Existing datasets include unannotated speech collections for speech representation learning, such as VoxLingua107 (Valk and Alumäe, 2020), VoxPopuli (Wang et al., 2021), JesusDramas, WikiTongues, and MMS ULAB v2 (Chen et al., 2024), which are mostly derived from social media, religious texts, or read speech for Standard German. Automatic speech recognition (ASR) datasets include FLEURS (Conneau et al., 2022), and dataset from initiatives like Mozilla Common Voice provide more structured recordings, while TTS datasets such as MLS (Pratap et al., 2020) (also useful for ASR), FLEURS-R (Ma et al., 2024), Thorsten-Voice Dataset (Müller, 2024), and CML-TTS (Oliveira et al., 2023), created from MLS, include high-quality recordings of standard German.

In recent years, some efforts have targeted German dialects for ASR and TTS, covering varieties such as Swiss German (Plüss et al.), Viennese (Schabus et al., 2013; Pucher et al., 2010), non-Vienna Austrian (Pucher et al., 2017), Luxembourgish (Steiner et al., 2017), and three dialect groups spoken in Southeast Germany (Franconian, Bavarian, Alemannic) (Blaschke et al., 2025). However, several dialects, such as the Saarbrücken dialect, remain underrepresented, and to the best of our knowledge only a small corpus is available in Bible MMS (Lux et al., 2024). Furthermore, a recent survey (Blaschke et al., 2024) showed that German speakers are interested in language technologies that can process dialectal (audio) input, and creating and evaluating such systems requires developing speech corpora for underrepresented dialects.

To address these gaps, we introduce **Saar-Voice**, a multi-speaker speech corpus for the Saarbrücken dialect, providing high-quality recordings suitable for training and evaluating modern TTS systems.

Speech Corpora for Dialects and Low-Resource Languages Beyond German-specific efforts, the creation of speech corpora for dialectal and low-resource settings has been widely explored across languages (Zampieri et al., 2020; Ramponi, 2024; Lent et al., 2022; Guellil et al., 2021; Alabi et al., 2025c). Studies on dialectal variants across a range of languages, such as Arabic (Malmasi and Zampieri, 2016; Djanibekov et al., 2025; Talafha et al., 2025), English (Ahamad et al., 2020; Xiao et al., 2023; Olatunji et al., 2023), and African languages (Ahia et al., 2024; Emezue et al., 2024), highlight the importance of accounting for regional and social variation when developing speech datasets, as models trained on standardized varieties often fail to generalize to non-standard speech (Diab, 2016; Aji et al., 2022; Ahia et al., 2024; Alabi et al., 2025b).

In parallel, a substantial body of research has focused on speech data collection for low-resource and underrepresented languages. These settings are typically characterized by limited availability of speakers, scarce linguistic resources, and, in some cases, the absence of standardized orthography (Blaschke et al., 2024). As a result, corpus construction in such contexts often relies on adaptable methodologies, including crowd-sourced data collection, community-driven recording initiatives, and

Speakers	
Total speakers	9
Gender ratio	4F / 5M
Speech	
Recorded sentences	4,871
Total duration (hh:mm:ss)	05:54:48
Avg. duration per sent. (sec)	4.37
Min duration of sent. (sec)	1.59
Max duration of sent. (sec)	14.02

Table 2: Saar-Voice Corpus Statistics.

semi-supervised or lightweight annotation strategies (Emezue et al., 2024; Olatunji et al., 2023).

Despite progress, many speech datasets focus only on common languages and pay little attention to dialect differences. In addition, datasets for low-resource languages often lack consistent annotation or enough speakers, which limits their usefulness for training and evaluating models. To address these challenges, we introduce a carefully curated speech corpus of a German dialect, with nine speakers and about six hours of recordings. The dataset is designed to provide a compact but high-quality resource for studying non-standard German speech and for training and benchmarking speech models, especially TTS, in low-resource dialect settings.

3. Corpus Design

This section describes the methodology used to create the Saar-Voice corpus.

3.1. Text Collection

Digitization of Printed Books For this study, we digitalized four books available in print at the Saarland University library. The books are "An da Saar gefonn: volkstümliche Gedichte in Saarbrücker Mundart" (Jungmann, 1993), "Geschaffd - Geläabd: Mundarttexte" (Fox, 1994), "Saa, was de willschd: Mundart-Kolumnen" (Fox, 2012) and "Was wääs dann isch...?!" (Eckert, 1995). These texts were scanned and digitized using an online OCR software[2]. This OCR software was chosen in particular as it was one of the few softwares that had close to no issue recognizing the special characters "á" and "ò", which caused great problems with other softwares and were usually misclassified as "ä" and "ö", respectively. Remaining errors introduced by the OCR process were removed through manual review by a native speaker of the dialect, while also cross-referencing the original source texts. 5,833 sentences were collected this way, amounting for 66.6% of the full dataset.

[2]https://ocr.ac/de

Locally Sourced Texts 2,838 sentences, accounting for 32.4% of the total data, were collected from internally available texts written by authors from the local community. These texts were already available digitally and only checked for possible spelling errors by a native speaker.

Localized Translations Lastly, we sampled 101 German sentences from the MASSIVE dataset (FitzGerald et al., 2023). These sentences were manually translated with the support of the dictionary of Braun and Mangold (1984). After translation, the data was localized by replacing entities with localized entities, such as locations and personal names. These localized entities were manually inserted to further emphasize dialect-specific variation. Localized location names were taken at random from Braun (1991), as in Example 1. Additionally, numbers were spelled out in dialect orthography rather than given as numerals, as in Example 2, to ensure consistent pronunciation across speakers.

(1) **German Original**: Bitte plane ein Treffen mit Petra in Wiesbaden am Mittag.
German Original (Entities Replaced): Bitte plane ein Treffen mit Anna in Oberbexbach am Mittag.
Translated: Bidde plaan e Dreffe midd Anna in Oberbexbach am Middaach.
Localized: Bidde plaan e Dreffe midd Anna in Owwerbeddschbach am Middaach.

(2) **German Original**: Erinnere mich an Lauras Geburtstag 24 Stunden vorher.
German Original (Entities Replaced): Erinnere mich an Melanies Geburtstag 19 Stunden vorher.
Translated: Erinner misch an Melanies Geburdsdaa 19 Schdunne vòòrhäär.
Localized: Erinner misch an Melanies Geburdsdaa neindsehn Schdunne vòòrhäär.

An overview over text statistics for each of the collected resources can be found in Table 1. This includes an estimation of the domains each of the resources cover, as annotated by a native speaker. Most resources are collections of texts rather than a single running texts. Therefore, for resources that are associated with multiple domains, domains are reported in descending order of frequency.

3.2. Speaker Recruitment

A total of nine participants were recruited using convenience sampling from people known to the research team. The speaker group consisted of four female and five male speakers. Age of participants was collected in categorical ranges. The largest age group was 26-30 years (n=3), followed by 31-35 years (n=2). The remaining participants were distributed across the ranges 18-25, 51-55, 56-60 and 61-65 (n=1 each).

Participants were either native speakers of a Rhine Franconian dialect, or native speakers of a closely related regional dialect with high familiarity and regular, long-term exposure to the target dialect. Self-identified dialect labels included: Saarländisch / Saarländischer Dialekt (*Saarland Dialect*; n=5), Moselfränkisch (*Moselle Franconian*, n=1), Rheinfränkisch (*Rhine Franconian* n=1), Rhein-Moselfränkischer Grenzdialekt (*Rhine-Moselle Franconian Border Dialect*, n=1), and Platt (*local term for the dialect*, n=1).

All speakers reported speaking the dialect daily (n=6) or at least multiple times a week (n=3) in the present time, and the majority (n=7) reports having spoken mainly the dialect during their childhood.

All speakers are also (native) speakers of Standard German and were "alphabetized" with Standard German spelling.

3.3. Recording Setup

Participants were invited to the departmental soundproof recording booth for audio recordings. The texts were presented to the speakers sentence by sentence on a 21.5-inch Full HD monitor. Audio was recorded at a sampling rate of 44.1kHz using a *DAP 2011* microphone, a high-quality directional microphone, using the SpeechRecorder Software (Draxler and Jänsch, 2004). Each recording session took around 1 hour, during which, depending on the participant, 200-300 sentences were recorded in batches of 100.

The speakers controlled their own pace of sentence presentation and recording by mouse clicks on a specified icon. If the speaker felt unsure or had a slip of the tongue s/he had the option for a new recording. Also, the recording supervisor (the first author) had the possibility to ask the speaker to restart the recording of a sentence.

Sentences presented to the speakers were not sampled independently, but derived from a longer, coherent text corpus. The entire textual corpus was concatenated, segmented into sentences and then grouped into chunks of 100 consecutive sentences. For each session, such chunks were randomly selected. This semi-randomized procedure ensured topical and lexical continuity within one session, while still maintaining coverage across the corpus and across domains.

Recording continuous passages like this offers several advantages. Besides keeping the speaker engaged in a cohesive story, it supports a more natural prosody and fluency, leading to consistent pronunciation of recurring and possibly unfamiliar lexical items, and enabling the resulting book subset to be used for long-context TTS experiments.

4. Saar-Voice Corpus

This section presents an overview of the created corpus.

4.1. Data Statistics

Table 2 summarizes the dataset statistics of Saar-Voice. The dataset consists of approximately six hours of speech, comprising 4,871 utterances from nine participants. The average length of the recorded sentence is about four seconds, with minimum and maximum durations of two and 14 seconds, respectively; these are moderately short but of good quality.

4.2. Phoneme Coverage

We applied Epitran (Mortensen et al., 2018) a multilingual grapheme-to-phoneme (G2P) model on the recorded text in Saar-Voice. Our analysis shows that the dataset contains 38 distinct phonemes, which lies in the typical range of phonemes estimated for Standard German (Kohler, 1990) and also the Saarbrücken dialect (Steitz, 1981) and (Braun and Mangold, 1984), if the vowels of German loan words are included. The most common phoneme is /d/, likely due to the voicing of many occurrences of consonants in the dialect that are typically unvoiced in Standard German (e.g., /t/ $\rightarrow$ /d/). It occurs 13,963 times within the speech corpus, and is followed by /n/ (11,284 occurrences), /s/ (10,963), /ʁ/ and /e/ (both 10,547 occurrences).

The least common phoneme is /ø/, occurring only 4 times in the entire speech corpus, followed directly by /œ/ (6 occurrences). Both are realizations of the Umlaut "ö", which is rather rare in the text itself, appearing only 10 times in total. They are followed by /ʏ/ (13 occurrences), /y/ (26 occurrences) and /ʒ/ (64 occurrences). All rounded front vowels /œ, ʏ, y/ can be regarded as phonemes from Standard German loan words.

The Epitran G2P model is multilingual and designed to cover standard German rather than the Saarbrücken dialect. While generally effective for estimating the phoneme coverage of this dialect corpus, the implications of this mismatch are discussed in Section 5.

4.3. Speaker Variability

Analysis of speaker variability shows that mean F0 (pitch) differs by gender. For male speakers, the mean F0 ranges from 115 Hz to 171 Hz (with a median of 107-164 Hz), while for female speakers it ranges from 201 Hz to 251 Hz (with a median of 196-249 Hz). These statistics indicate moderate inter-speaker variability in both pitch and tempo, reflecting natural differences in speech across male

Speaker	Gender	Total audio (hh:mm:ss)	# sentences	Mean F0	Median F0	F0 range (5–95%)	W/s
P01	F	00:23:17	301	203	200	141-278	1.8249
P02	F	01:02:28	1075	251	249	203-312	2.4166
P03	M	01:15:37	899	115	107	80-176	1.7352
P04	M	00:27:56	400	143	140	100-196	2.2056
P05	F	00:27:18	399	213	212	175-256	2.0203
P06	M	00:18:43	300	119	115	84-160	2.3334
P07	M	00:53:21	597	171	164	102-258	1.5016
P08	F	00:44:45	600	201	196	156-265	1.7920
P09	M	00:21:22	300	147	141	101-203	1.8053

Table 3: Speaker-level acoustic statistics. F0 in Hz. W/s = words per second.

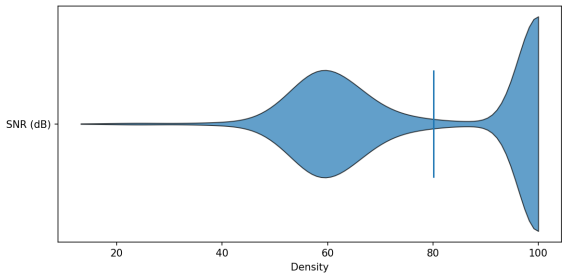


Figure 2: SNR distribution on Saar-Voice.

and female voices. We also observe speech rates ranging from 1.5 to 2.4 words per second across both genders, with values evenly distributed within each gender. Table 3 provides a detailed breakdown for all speakers.

4.4. Audio Quality

We evaluate the signal-to-noise ratio (SNR) for all the recorded audio samples in Saar-Voice using the WADA-SNR (Kim and Stern, 2008) algorithm and plot the distribution in Figure 2. The results show that most samples fall within the clean audio threshold, with SNR values ranging from 20 to 99, confirming the high overall quality of the Saar-Voice dataset.

5. Issues

In this section, we discuss the practical challenges and issues encountered during the corpus creation process.

5.1. Data-Specific Issues

A central challenge of building a corpus for a low-resource dialect like the variant of Rhine Franconian discussed in this project is the absence of standardized orthography for the target dialect. Unlike Standard German, the dialects of the Saarland do not have any official spelling conventions or vocabulary lists. Although dialect dictionaries like Braun and Mangold (1984) can serve as guidance, they do not provide an official, normative reference. Orthographic variation is thus unavoidable and inherent to the dialect. It further affects vocabulary size estimates, as orthographic variants of the same lexical item are counted as distinct types.

Ignoring such variation or attempting to impose a standardized orthography would in many cases even go against what native speakers actually desire: Blaschke et al. (2024) show that 65% of their participating speakers of German dialects explicitly oppose the introduction of a standardized orthography. Our own questionnaire showed no clear consensus regarding standardization. Taken together, these results suggest that there is no stable community-wide preference for standardization, and the more representative survey by Blaschke et al. (2024) indicates the opposite. Any attempt at standardization would risk privileging one subset of speaker preference over others, potentially misrepresenting actual usage practices, and harming user trust.

The choice of OCR software presented an additional technical challenge. Most available systems, such as Python's *tesseract* library, are optimized and effective for Standard German, but showed a substantial difficulty processing dialect-specific characters such as “á” or “ò”, as well as recognizing entire character strings which are not part of Standard German's inventory. This resulted in recognition errors and, even with the final choice of OCR software, required manual corrections, introducing additional processing effort and potential inconsistencies.

Further issues with available technologies arise in the lack of dedicated tokenization tools for the dialect. This led to being able to provide only an approximate vocabulary segmentation by whitespace tokenization. Without dialect-specific morphological or tokenization resources, which are hard to create due to data sparseness, word boundary detection may be imperfect. This in turn affects vocabulary size estimation and lexical frequency calculations.

Beyond orthography issues, the composition of the corpus itself raises questions of linguistic representativeness. The genre of poetry makes up

over one third (35.9%) of the dataset. This may, due to the genre's metrical and rhyming conventions, introduce lexical and especially prosodic and suprasegmental patterns that are unlikely to reflect the natural spoken dialect. More broadly, all textual resources, including prose and localized translations, represents the written register language, which inherently diverges from spontaneous speech. Nonetheless, it accurately represents the distribution of written data within this specific dialectal landscape. Additionally, the dataset lacks spontaneous speech. While this is less critical for the dataset's primary intended use case of TTS synthesis, it may limit its application to downstream tasks such as ASR. Future work should therefore focus on expanding the corpus with transcribed spontaneous speech to improve linguistic representativeness and genre coverage, and broaden its utility.

5.2. Speaker-Specific Issues

Although all speakers were carefully selected as Rhine Franconian speakers, variation within the speaker pool naturally remains inevitable. A variety of self-assessment data from the provided questionnaire indicates high overall proficiency / fluency, but variance appears in dialect usage patterns. Only one speaker reports speaking predominantly exclusively dialect with little to no Standard German influences, while most described their speech as a mixture of dialect and standard language. Some even indicated that they predominantly speak Standard German.

These differences, as well as considering small regional differences in the dialect, suggest that the corpus reflects not a uniform dialect realization, but rather a spectrum of usage patterns. This introduces additional variability for downstream modeling tasks.

5.3. G2P-Specific Issues

Beyond general data limitations, several issues emerged specifically during the G2P conversion process. Representative examples for each of the following issues can be found in Table 4.

Firstly, the most frequent detected IPA symbol in the output is the length mark /ː/, indicating that the model predicts a high number of lengthened vowels. However, through manual inspection, it becomes clear that vowel length is often assigned incorrectly. Words ending in the letter "e", are often transcribed to end in /eː/, which is incorrect. This mistake stands out in particular as the correct transcription would generally be /ə/, just like in Standard German. Also, it appears to misclassify double vowels, which are common in dialectal writing, as a sequence of two lengthened vowels.

Additionally, a common phenomenon in dialectal writing is the occurrence of a doubling of the letter "d", which is rather rare in Standard German and instead often corresponds to the Standard German "tt". This double consonant is misinterpreted by the model as /td/ on several occasions, which reflects an incorrect segmentation.

Lastly, the Epitran (Mortensen et al., 2018) G2P model generated some characters which are impossible in the German phoneme inventory, such as /m̀/, showing expected issues with processing the special characters "á" and "ò".

6. Dialectal TTS Modeling

The primary motivation behind creating Saar-Voice is to enable the development and evaluation of multi-speaker TTS systems. With recent advances in multilingual and multi-speaker (zero-shot) TTS models, such as XTTS (Casanova et al., 2024) and ZMM-TTS (Gong et al., 2024), it is now possible to investigate how well multilingually pretrained systems can generalize to closely related dialectal varieties, even when only limited data are available.

Both models include German in their pretraining, making adaptation to Saarbrücken dialect potentially feasible. However, recent research on low-resource Bildts (a Dutch variety) (Do et al., 2025) observed speaker–language entanglement when adapting ZMM-TTS for unseen speakers (zero-shot), and for seen speakers, there were also multiple cases of mispronunciation (Alabi et al., 2025a). In contrast, Pine (2025) shows that combining Dutch and Bildts and using the StyleTTS (Li et al., 2023) architecture is sufficient for the same task. Hence, our dataset will help to better understand how multilingually pretrained TTS models and architectures generalize to closely related dialectal varieties.

Such evaluation of zero-shot and fine-tuned TTS models using the dataset are a goal of future work.

7. Conclusion

This paper presented the creation of a multi-speaker corpus for a low-resource German dialect. The dataset was constructed using partially OCR-based text extraction and manual normalization, and speech was produced by a variety of speakers. The resulting corpus provides both textual and phonetic representation suitable for speech technology research, with the option to expand the quantity of speech data. The challenges faced in the creation of the corpus reflect broader issues in low-resource dialect modeling. While they may introduce variability, they capture what authentic dialect usage looks like in native speakers.

Word	English	Epitran	Ours
berechne	calculate (Imp.)	bəʀɛxneː	bəʀɛxnə
bidde	please	bɪtdeː	bɪdə
Midde	middle	mɪtdeː	mɪdə
odder	or	ɔtdəʀ	ɔdɐ
buuch	book (Imp.)	buːuːx	buːx
ääner	one	æːæːnəʀ	æːnɐ
òòmens	in the evening	oːòːm̀əns	ɔːmns
Schdigg	piece	ʃdɪçk	ʃdɪg

Table 4: Examples of G2P Errors. Corrections of transcriptions are estimates by a native speaker.

Empirical validation of the dataset through downstream experiments remains an important next step. TTS synthesis experiments using the corpus are left for future work. Through possible zero-shot inference, model fine-tuning and training approaches, further evidence of the dataset's utility for speech technology applications in this low-resource dialect setting may be provided.

Beyond the dataset creation and proposing use cases, we aim to emphasize the necessity of collaborating with the speaker community. Without direct engagement with the dialect writers' scene and the speakers, the amount of texts would have been dramatically reduced. The existence of this corpus is therefore closely tied to the active dialect-writing community.

In dialectal research and Natural Language Processing (NLP) as a whole, methodological decisions should not be guided solely by technical feasibility or research interest, but also by the expectations and preferences of the speaker community itself. Especially in low-resource and non-standard contexts, technological developments interacts directly with questions of representation, identity, language maintenance and culture.

Our questionnaire, following Blaschke et al. (2024), asked speakers directly whether they would welcome more digital applications supporting dialect spoken in the Saarland. The large majority of our participants expressed agreement (strong agreement: n=6; agreement: n=2; neutral: n=1). Similarly, most respondents agree that language technologies can contribute to dialect preservation (strong agreement: n=5; agreement: n=3; neutral: n=1). Despite the limited sample size and biased participant pool, these responses suggest that technological support may be perceived rather positively within the community.

Thus, we consider it an important detail that dialect NLP projects have a contact to the various communities using dialects for getting

- access to written texts and spoken materials,
- recommendations for book authors and relevant sources,
- insights into possible or prominent spelling patterns / conventions and variations,
- feedback on acceptable and desirable technological applications.

Community engagement does and should not only serve as a practical resource for data collection, but also as a means of aligning technological development with actual users.

8. Acknowledgments

We would like to thank all speakers who contributed their voices to this corpus, without whom this resource would not have been possible. Many thanks to the anonymous reviewers for their time and efforts, and for the extensive feedback and questions. Thank you also to our colleague Çağla Kints for carefully proof-reading this paper. Jesujoba Alabi was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

9. Supplementary Materials

9.1. Ethical Considerations

Informed consent. Participants signed a consent from in which they agreed to the process of the recording as well as the anonymized storage, sharing and processing of the recorded data. No personally identifying information of speakers was recorded, and all data is exclusively linked to an anonymized, randomized speaker ID. Participants were also informed that they could withdraw their consent at any time.

Data protection. All recorded data is stored in an anonymized form, only linking to the randomized speaker ID, on secure university servers.

Voluntariness. Participation was completely voluntary and occurred without compensation. No material incentives were offered to participants, and participation had no academic or professional consequences in any way.

Participant well-being. Great care was taken of keeping the duration of recording sessions to

around one hour, though participants were free to end their sessions early if they wished. Breaks were always possible, but taken at least after the recording of 100 sentences. Participants had the option to prepare for the recording, as they received the material to be read during their session digitally around one week in advance.

Community sensitivity. The dataset represents just a subset of the regional dialects of the Saarland. The process of curating the data involved no evaluation of dialectal "correctness", including no aim to standardize or change dialect varieties. The textual content of the dataset reflects the original authors' perspectives and was not modified or filtered to promote or follow any particular set of views.

10. Bibliographical References

Astrid Adler and Karolina Hansen. 2022. Dialekt und Beruf: neue Daten zu Dialekten in Deutschland. Sprache in Zahlen: Folge 7. *Sprachreport*, 38(3):28–33.

Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. 2020. AccentDB: A database of non-native English accents to assist neural speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5351–5358, Marseille, France. European Language Resources Association.

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. Voices unheard: NLP resources and models for Yorùbá regional dialects. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4392–4409, Miami, Florida, USA. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Jesujoba O. Alabi, Cheng Gong, Erica Cooper, Yu Jiang, Dietrich Klakow, and Junichi Yamagishi. 2025a. Submission from ZMM-TTS for Blizzard Challenge 2025. In *The Blizzard Challenge 2025*, pages 31–36.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2025b. AfriHuBERT: A self-supervised speech representation model for African languages. In *Interspeech 2025*, pages 4023–4027.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025c. Charting the landscape of African NLP: Mapping progress and shaping the road ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? A survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Minh Duc Bui, Carolin Holtermann, Valentin Hofmann, Anne Lauscher, and Katharina von der Wense. 2025. Large language models discriminate against speakers of German dialects. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8212–8240, Suzhou, China. Association for Computational Linguistics.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

Mona Diab. 2016. Processing dialectal Arabic: Exploiting variability and similarity to overcome challenges and discover opportunities. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, page 42, Osaka, Japan. The COLING 2016 Organizing Committee.

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. Dialectal coverage and generalization in Arabic speech recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502, Vienna, Austria. Association for Computational Linguistics.

Phat Do, Matt Coler, Jelske Dijkstra, Igor Marchenko, Vass Verkhodanova, and Sebastien Le Maguer. 2025. The Blizzard Challenge 2025. In *The Blizzard Challenge 2025*, pages 1–17.

Christoph Draxler and Klaus Jänsch. 2004. SpeechRecorder - A universal platform independent multi-channel audio recording software. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Chris Chinenye Emezue, Ifeoma Okoh, Chinedu Emmanuel Mbonu, Chiamaka Chukwuneke, Daisy Monika Lal, Ignatius Ezeani, Paul Rayson, Ijemma Onwuzulike, Chukwuma Onyebuchi Okeke, Gerald Okey Nweya, Bright Ikechukwu Ogbonna, Chukwuebuka Uchenna Oraegbunam, Esther Chidinma Awo-Ndubuisi, and Akudo Amarachukwu Osuagwu. 2024. The IgboAPI dataset: Empowering Igbo language technologies through multi-dialectal enrichment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15932–15941, Torino, Italia. ELRA and ICCL.

Cheng Gong, Xin Wang, Erica Cooper, Dan Wells, Longbiao Wang, Jianwu Dang, Korin Richmond, and Junichi Yamagishi. 2024. Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:4036–4051.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Chanwoo Kim and Richard M. Stern. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Interspeech 2008*, pages 2598–2601.

Klaus Kohler. 1990. German. *Journal of the International Phonetic Association*, 20(1):48–50.

Xaver Maria Krückl, Verena Blaschke, and Barbara Plank. 2025. Improving dialectal slot and intent detection with auxiliary tasks: A multi-dialectal Bavarian case study. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 128–146, Abu Dhabi, UAE. Association for Computational Linguistics.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc.

Shervin Malmasi and Marcos Zampieri. 2016. Arabic dialect identification in speech transcripts. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 106–113, Osaka, Japan. The COLING 2016 Organizing Committee.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Alberto Muñoz-Ortiz, Verena Blaschke, and Barbara Plank. 2025. Evaluating pixel language models on non-standardized languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6412–6419, Abu Dhabi, UAE. Association for Computational Linguistics.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Aidan Pine. 2025. The NRCC Submission to the Blizzard Challenge 2025. In *The Blizzard Challenge 2025*, pages 24–30.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, AbdelRahim A. Elmadany, Abdurrahman Juma, Amirbek Djanibekov, Chiyu Zhang, Hamad Alshehhi, Hanan Aldarmaki, Mustafa

Jarrar, Nizar Habash, and Muhammad Abdul-Mageed. 2025. NADI 2025: The first multidialectal Arabic speech processing shared task. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 720–733, Suzhou, China. Association for Computational Linguistics.

Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen English dialects. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7857–7870, Singapore. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

11. Language Resource References

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, pages 913–917.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Edith Braun. 1991. *Necknamen der Saar und drum herum*. Hempel-Verlag, Lebach.

Edith Braun and Max Mangold. 1984. *Saarbrücker Wörterbuch*. Saarbrücker Druckerei und Verlag.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Georg Drenda. 2008. *Kleiner linksrheinischer Dialektatlas. Sprache in Rheinland-Pfalz und im Saarland*. Franz Steiner Verlag.

Peter Eckert. 1995. *Was wääs dann isch ...?! : hundertzwanzigmal Mundart von der Saar*. KleinstLiteratur aus der edition händmäid. KleinstVerl. Die Kiste, Wadgassen-Differten.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECT-BENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Georg Fox. 1994. *Geschaffd - gelääbd: saarländische Mundarttexte*. cjm-Verlag, Speyer.

Georg Fox. 2012. *Saa, was de willschd!: Mundart-Kolumnen*. PVS-Edition, Heusweiler.

Kurt Jungmann, editor. 1993. *An da Saar gefonn: volkstümliche Gedichte in Saarbrücker Mundart*. Westwind-Verlag, Saarbrücken.

Florian Lux, Sarina Meyer, Lyonel Behringer, Frank Zalkow, Phat Do, Matt Coler, Emanuël A. P. Habets, and Ngoc Thang Vu. 2024. Meta Learning Text-to-Speech Synthesis in over 7000 Languages. In *Interspeech 2024*, pages 4958–4962.

Min Ma, Yuma Koizumi, Shigeki Karita, Heiga Zen, Jason Riesa, Haruko Ishikawa, and Michiel Bacchiani. 2024. FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks. In *Interspeech 2024*, pages 1835–1839.

Thorsten Müller. 2024. Tv-44khz-full (revision ff427ec).

Frederico S. Oliveira, Edresson Casanova, Arnaldo Candido Junior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts: A multilingual dataset for speech synthesis in low-resource languages. In *Text, Speech, and Dialogue*, pages 188–199, Cham. Springer Nature Switzerland.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. Swiss parliaments corpus, an automatically aligned Swiss German speech to standard German text corpus.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Michael Pucher, Carina Lozo, and Sylvia Moosmüller. 2017. Phone mapping and prosodic transfer in speech synthesis of similar dialect pairs. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pages 180–185. TUDpress, Dresden.

Michael Pucher, Friedrich Neubarth, Volker Strom, Sylvia Moosmüller, Gregor Hofer, Christian Kranzler, Gudrun Schuchmann, and Dietmar Schabus. 2010. Resources for speech synthesis of viennese varieties. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2013. Joint audiovisual hidden semi-markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):336–347.

Ingmar Steiner, Sébastien Le Maguer, Judith Manzoni, Peter Gilles, and Jürgen Trouvain. 2017. Developing new language tools for MaryTTS: the case of Luxembourgish. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, pages 186–192. TUDpress, Dresden.

Lothar Steitz. 1981. *Grammatik der Saarbrücker Mundart*. Saarbrücker Druckerei und Verlag.

Jörgen Valk and Tanel Alumäe. 2020. Voxlingua107: A dataset for spoken language recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

MD_NLP: Reconstructing an Australian English Heritage Dialect Corpus from the Mitchell-Delbridge Recordings through LLM-Assisted Speaker Attribution

Steven Coats
English, Faculty of Humanities, University of Oulu
90014 University of Oulu, Finland
steven.coats@oulu.fi

Abstract

We present MD_NLP, a discourse-annotated and georeferenced corpus derived from the Mitchell–Delbridge (MD) recordings, a foundational archive of mid-20th-century Australian English. The corpus comprises word-aligned narrative recordings from 7,735 secondary school pupils across 327 locations, enriched with structured sociodemographic metadata (e.g., sex, birthplace, parental background) and geocoded institutional coordinates. The narratives were reconstructed from archival audio using an integrated pipeline combining WhisperX-based automatic speech recognition, neural speaker diarization, LLM-assisted discourse-role correction, and Montreal Forced Aligner (MFA) boundary refinement. Evaluation on manually annotated data shows that incorporating an LLM-based reasoning step improves turn-level speaker-role attribution from 62.70% (acoustic diarization alone) to 95.68%. Unlike prior uses of the MD archive, which focused on controlled sentence materials, MD_NLP makes the spontaneous narrative component accessible for large-scale analysis. The resulting resource supports research on regional and socially conditioned variation, discourse structure, and corpus phonetics in Australian English. The proposed architecture is directly transferable to other legacy dialect archives, providing a practical pathway for transforming interview-based recordings into temporally aligned, speaker-consistent corpora.

Keywords: dialect corpora, Australian English, speaker diarization, forced alignment, discourse annotation, large language models

1. Introduction

Automatic Speech Recognition (ASR) systems are increasingly used to generate transcripts for dialectal speech in corpus phonetics and variationist linguistics (Ahn et al., 2023; Coto-Solano et al., 2021; Coats et al., 2025a; Ljubešić et al., 2024). However, the transformation of raw ASR output into research-ready dialect resources remains underdeveloped. For many legacy collections, usable corpora require not only transcription, but also reliable speaker attribution, accurate word-level timing, and discourse-structural annotation.

This paper introduces MD_NLP, a new discourse-annotated and word-aligned version of the Mitchell and Delbridge (MD) recordings (Mitchell and Delbridge, 1998). The MD collection comprises recordings of 7,735 Australian secondary school pupils from 327 schools across Australia, recorded in 1959/60 and digitized at the University of Sydney in 1998. The collection constitutes one of the largest geographically distributed archives of mid-20th-century Australian English (AusE). Despite its importance for the study of AusE, including regional and social variation, the narrative component of the corpus has remained largely inaccessible for large-scale analysis due to the absence of transcripts and the variable acoustic quality of the archival recordings.

MD_NLP addresses this gap by providing speaker-labeled, word-aligned transcripts of the MD narratives with precise temporal metadata. In addition, the corpus integrates structured sociodemographic and geographic information derived from the original Mitchell–Delbridge archival records, including sex, birthplace, parental background, school location, and geocoded coordinates. The resulting resource supports not only phonetic investigation, but also lexical, grammatical, discourse-level, and spatial sociolinguistic analyses.

Over the past two decades, research on Australian English (AusE) has increasingly examined regional phonetic differentiation (Schmidt et al., 2021; Loakes et al., 2017; Coats et al., 2025b; Cox and Palethorpe, 2019). Yet large-scale, geographically distributed spontaneous speech data remain scarce. By making nationally distributed narrative recordings from 1959/60 accessible in a temporally aligned and georeferenced format, MD_NLP enables systematic investigation of regional and socially conditioned variation in historical AusE.

To make this historical, geographically distributed material computationally accessible, we developed an integrated processing pipeline. The corpus was constructed using a hybrid pipeline that combines WhisperX-based automatic speech recognition, neural speaker diarization, LLM-assisted

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 24–32
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

discourse role inference and diarization correction (Gemini), and high-precision forced alignment using the Montreal Forced Aligner (MFA), resulting in automated segmentation into short, role-consistent narrative units.

Building on recent work that leverages Large Language Models (LLMs) to refine speaker attribution in ASR transcripts (Wang et al., 2024; Cheng et al., 2025), our approach addresses two core challenges in dialect resource construction. First, neural diarization models based solely on acoustic information, such as Pyannote (Bredin, 2023; Plaquet and Bredin, 2023), frequently misattribute interviewer speech or collapse speakers in archival classroom-style recordings. We incorporate a reasoning-oriented LLM step to correct these errors by exploiting interactional structure (e.g., question–answer sequences), substantially improving speaker-role consistency. Second, we integrate high-precision word-level alignment via the Montreal Forced Aligner (McAuliffe et al., 2017), producing temporal annotations suitable for corpus-phonetic and other variationist analyses that require fine-grained timing information.

MD_NLP provides both a research-ready resource for Australian English dialect research and a transferable pipeline for converting legacy dialect recordings into temporally precise, discourse-annotated corpora.

2. Previous Work

2.1. Use of the MD Recordings in Australian English Research

The MD archive comprises read word lists, two controlled sentence passages, and spontaneous narrative recordings. To date, research has focused exclusively on the controlled sentence materials, while the narrative component has not been systematically analyzed.

Cox et al. (2014) examined the development of /i:/ in the speech of young female Sydneysiders by comparing 168 realizations in the word *speed* from the MD read sentence with an equivalent number of recordings made in the 2000s. The authors found that while the overall quality of the vowel had not changed substantially over time, differences between "broad" and "cultivated" realizations appeared to have lessened. In Cox et al. (2024), static and dynamic formant measurements were taken for ten monophthongs extracted from the MD sentence materials and compared with equivalent recordings from 1989/90 and the 2000s/2010s, shedding light on the diachronic development of Australian English vowels.

The limited use of the narrative material is largely attributable to the absence of transcripts. Manual transcription of thousands of archival recordings is time-consuming and costly, and the variable acoustic quality of the MD recordings poses additional challenges for automated processing.

2.2. LLM-Assisted Diarization and Speaker Attribution

Recent work has explored the use of Large Language Models (LLMs) to improve speaker attribution in ASR transcripts. Wang et al. (2024) introduced DiarizationLM, which leverages semantic and contextual reasoning to correct diarization errors through textual post-processing. Similarly, Cheng et al. (2025) demonstrated that LLM-based approaches can improve speaker consistency by exploiting discourse structure beyond purely acoustic cues.

These studies highlight limitations of diarization systems based solely on acoustic information. In interview-style or classroom recordings, short interviewer turns are frequently misattributed or merged with longer responses. Such errors are particularly problematic for sociolinguistic research, where reliable speaker separation and discourse-role identification are essential.

In addition, although modern ASR architectures such as Wav2Vec 2.0 (Baevski et al., 2020) and NeMo-based systems (Harper et al., 2023) provide word-level timestamps, these timings are primarily optimized for transcription accuracy rather than phonetic precision. For corpus-phonetic research, even small boundary inaccuracies can affect segmental measurements.[1]

Our work extends these lines of research by integrating LLM-based diarization correction with high-precision forced alignment using the Montreal Forced Aligner (MFA), producing temporally refined and speaker-consistent transcripts suitable for dialect and variationist analysis.

3. Source Material

The Mitchell and Delbridge recordings were made by schoolteacher volunteers in 1959 and 1960 in 327 locations across all Australian states and territories.

The recordings contain, for each informant, two to four components:

- The informant reading the words *beat, boot, say, so, high, and how*.
- The informant reading the sentence *Let's pick a good spot near the water and pass the morning surfing and relaxing in the sun*.

[1] See, e.g., MFA alignment benchmarks.

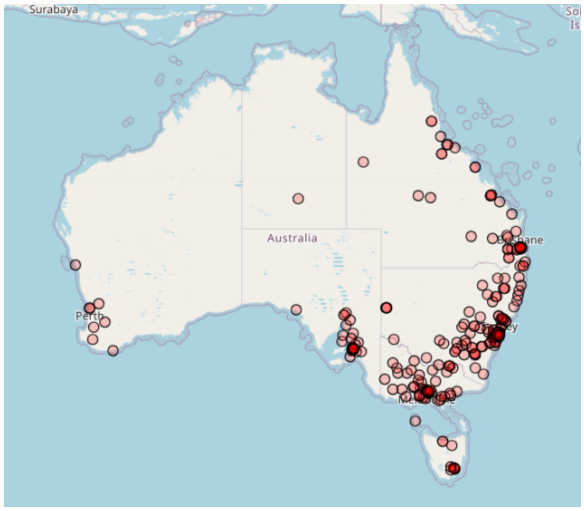


Figure 1: Sample locations for the MD narrative recordings

- The informant reading the sentence *The plane flew down low over the runway then increased speed and circled the aerodrome a second time*.
- The informant providing a short narrative on a topic of their choice. In many recordings, this takes the form of a brief interaction between the pupil and the recording teacher; in others, it consists primarily of a monologic narrative.

Recordings made in 1960 generally include all four components, whereas those made in 1959 may contain only the word list and the narrative.

MD_NLP comprises the narrative recordings from the archive. These data are particularly suitable for sociolinguistic and discourse-oriented research, as they contain spontaneous speech, question–answer sequences, and turn-taking phenomena, alongside detailed informant metadata including school, locality, and parental background. At the same time, the archival recording conditions and lack of transcripts present challenges for automated processing.

4. Pipeline Overview

Each input file is a single MD narrative WAV. We distinguish (i) WhisperX *turns* (initial segments), (ii) MFA *alignment windows* (short units used only for forced alignment), and (iii) released *dataset chunks* (final $\leq$12-token, role-consistent segments). Lexical forms always remain the WhisperX tokens; MFA is used only to refine timestamps when alignment succeeds.

Figure 2 outlines the processing pipeline. The system integrates ASR, diarization, LLM-based discourse reasoning, and forced alignment in a reconstruction-oriented workflow that preserves original tokenization while improving temporal precision.

4.1. Automatic Speech Recognition and Alignment

Preliminary transcripts were created using WhisperX (Bain et al., 2023), an implementation of Whisper (Radford et al., 2023) that utilizes the faster-whisper bindings (Klein and Ashraf, 2023) and Wav2Vec2 (Baevski et al., 2020) for word and phone alignment. We used the large-v3 Whisper model. WhisperX timestamps are treated as coarse alignment; we use MFA forced alignment to improve boundary precision for downstream phonetic and timing-sensitive analyses.

4.2. Speaker Diarization and Text Normalization

Speaker turns were identified with Pyannote.audio's speaker-diarization-community-1 model (Bredin, 2023; Plaquet and Bredin, 2023). The conditions of the MD recordings were highly variable in terms of recording equipment, microphone placement and recorded material. Essentially, each school decided how to record the short narrative, resulting in various outcomes for the resulting audio. In some recordings, a microphone was equally spaced between the teacher and the student informant, resulting in both speakers being audible. In others, the microphone was oriented towards the student. In these recordings, the speech of the interviewing teacher is often faint or inaudible. In some recordings, only the student speech has been recorded; presumably due to erasure of the segments containing the interviewer's speech.

We therefore adopted a flexible approach to recording diarization: pyannote's settings were defined to identify a minimum of 1 speaker and a maximum of two speakers. The output for most of the recordings was two speaker labels (SPEAKER_00 and SPEAKER_01). For the recordings with only faint/inaudible interviewer speech, and for those from which the interviewer's speech was removed, pyannote's output was the single speaker label SPEAKER_00.

The initial diarized transcripts were then normalized to ensure compatibility with MFA dictionaries, including expansion of symbols (e.g. "50%" → "50 percent") and removal of non-lexical punctuation. Each transcript was then converted to a TextGrid file with speaker tiers corresponding to the number of speakers detected in the previous step, resulting in an initial word-level transcript with timestamps.

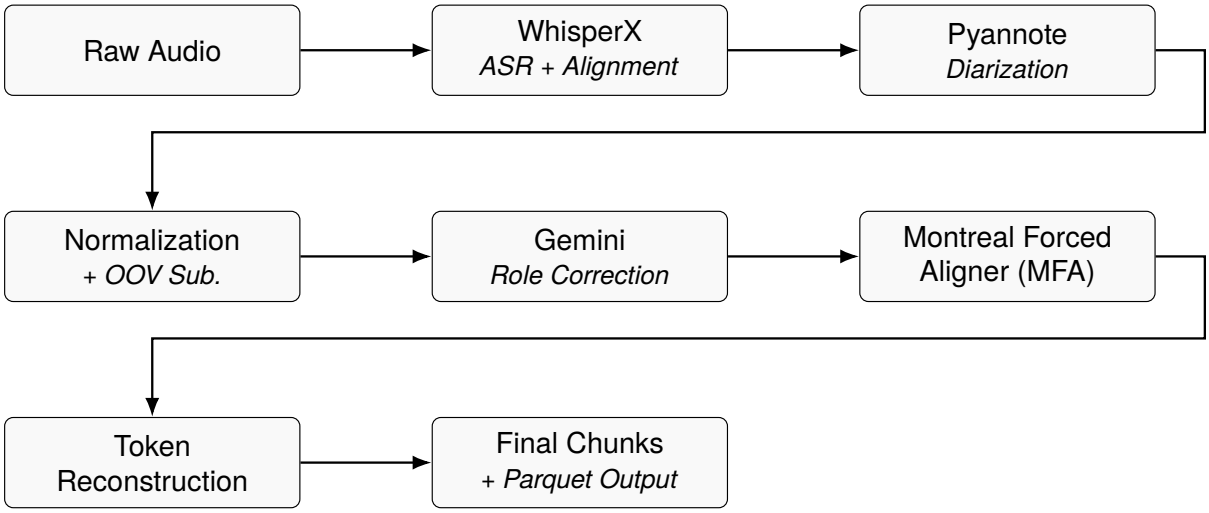


Figure 2: The integrated processing pipeline. Raw audio is transcribed with WhisperX, diarized with Pyannote, and corrected for discourse roles using an LLM. Word-level alignment is refined with MFA on short ($\leq$12-token) windows, after which original WhisperX tokens are reconstructed with selectively updated timestamps.

4.3. LLM-Assisted Discourse Role Inference

While the initial, acoustic-information-based diarization provided speaker separation, many of these labels were erroneous. In addition, the preliminary diarization could not determine discourse roles. In the MD setting, distinguishing between interviewer (Teacher) and respondent (Student) is essential for downstream analyses. We therefore applied an LLM-based reasoning step using Gemini (2.5-flash) to infer discourse roles from transcript context. The model was prompted with the following:

> Stage 1: If turns appear to be answers but no questions are present, set `inaudible_teacher=true` and label all turns STUDENT.
> Stage 2: Otherwise correct speaker attribution: Teachers ask questions; Students provide narrative/answers. Roles should be consistent per diarization label (e.g., `SPEAKER_00`).

A preliminary gate handles recordings where interviewer speech is faint/absent: if turns appear to be answers but no questions are present, all turns are labeled STUDENT to avoid forcing a spurious Teacher role. The LLM returned a structured JSON specifying corrected role assignments per turn. A qualitative example of this correction process is shown in Figure 3.

This step operationalized interpretive reasoning rather than purely acoustic modeling.

4.4. Forced Alignment and Token Reconstruction

Although WhisperX provides word-level timestamps, we further refined alignment using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), with version 3.0.0 of its UK English acoustic model, grapheme-to-phoneme model, and pronunciation dictionary. To increase alignment stability and avoid long-span decoding failures, alignment windows were capped at 12 tokens and further split on pauses > 0.8s or window duration > 8.0s to improve MFA stability.

Before MFA, out-of-vocabulary tokens were replaced with an in-vocabulary placeholder *for MFA input only*. In reconstruction, the original WhisperX token is always restored; if a token was placeholder-aligned (or missing), its original WhisperX timing is retained.

After MFA alignment, word intervals were reconstructed by mapping aligned tokens back onto the original Whisper word sequence. Tokens that were dropped or substituted for alignment were restored with their original Whisper timestamps, ensuring that no Whisper tokens were discarded, original lexical forms were preserved, and alignment precision was improved where MFA word intervals were available. Otherwise, Whisper timings were retained. This reconstruction step produced a temporally refined but lexically faithful word-level representation.

Baseline: WhisperX + Pyannote (Acoustic Diarization Only)

SPEAKER_01: Could you describe water skiing, what you do and what gear you need and where you do it?
SPEAKER_00: Well, the gear, first you need a boat, which seems essential.
SPEAKER_00: And the boat ranges from about, say, 13, 14 feet onwards.
SPEAKER_00: If you're going to learn to ski, you need two skis.
SPEAKER_00: How fast do you do it on that?
SPEAKER_00: Oh, about 30 is a pretty good speed because you can do certain different things on it when you're up.
SPEAKER_00: Do you hurt yourself if you come off?
SPEAKER_00: Oh, I've never hurt myself except once when I land on my head.

Full Pipeline: LLM-Assisted Role Attribution

Teacher: Could you describe water skiing, what you do and what gear you need and where you do it?
Student: Well, the gear, first you need a boat, which seems essential.
Student: And the boat ranges from about, say, 13, 14 feet onwards.
Student: If you're going to learn to ski, you need two skis.
Teacher: How fast do you do it on that?
Student: Oh, about 30 is a pretty good speed because you can do certain different things on it when you're up.
Teacher: Do you hurt yourself if you come off?
Student: Oh, I've never hurt myself except once when I land on my head.

Figure 3: Qualitative comparison of acoustic diarization and the LLM-assisted pipeline. The baseline system collapses interviewer and respondent speech into a single speaker label after the initial turn. Interviewer questions (highlighted in red) are incorrectly attributed to the same speaker as narrative responses. The LLM-assisted step recovers consistent Teacher–Student roles by exploiting interactional structure.

5. Dataset Structure

5.1. Segment-Level Structure

The final dataset is stored as Parquet files with corresponding audio segments, and is available as a Hugging Face dataset.[2] Each row corresponds to a short ($\leq$12-token), role-consistent segment and contains: (i) interviewee identifier, (ii) start and end time (in seconds), (iii) discourse role (Student/Teacher), (iv) transcript text, (v) word-level tokens with timestamps, and (vi) a path to the corresponding audio segment.

Word-level tokens are stored as structured objects containing lexical form, start time, and end time, preserving the original WhisperX tokenization while incorporating refined timestamps where available.

5.2. Sociodemographic and Geographic Metadata

In addition to transcript-level annotations, MD_NLP incorporates structured informant metadata derived from the original Mitchell–Delbridge archival records provided by the University of Sydney. The original metadata were distributed in heterogeneous spreadsheet formats and required normalization and cleaning prior to integration.

For each informant, the metadata include: identifier (ID#), state, town/city, institution (school), recording year, tape identifier, sex, birthplace, father's birthplace, father's occupation, and mother's birthplace. Institutional location fields were normalized and geocoded, resulting in latitude–longitude coordinates associated with each recording site.

These metadata are linked to each narrative segment via the informant identifier, enabling analyses that integrate discourse structure, phonetic timing, demographic attributes, and geographic location within a unified sociolinguistic resource.

5.3. Corpus Size

Table 1 summarizes corpus size and discourse distribution. The dataset comprises 214.14 hours of recordings, of which 137.95 hours are attributed to active speech following diarization and role assignment.

Table 1 shows the size of the dataset.

The corpus reflects broad geographic and gender representation across Australian states (Table 2).

[2] https://huggingface.co/datasets/stcoats/MD_NLP

Metric	Student	Teacher	Total
Recording Length (h)	–	–	214.14
Speech (h)	92.71	45.24	137.95
Turns	46,026	25,903	71,929
Word Count	1,155,994	635,862	1,791,856

Table 1: Comparative analysis of classroom discourse: Recording vs. active speech duration, turn frequency, and lexical volume by role.

State	Female	Male	Unk.
Australian Capital Territory	24	60	0
New South Wales	2,071	1,729	1
Queensland	617	617	0
Northern Territory	7	13	0
South Australia	402	424	1
Tasmania	173	116	0
Victoria	676	615	0
Western Australia	117	72	0

Table 2: Distribution of informants by state and recorded sex (archival metadata field).

6. Evaluation

6.1. Evaluation Setup

We randomly selected 10 narratives (approximately 30 minutes of speech; 185 WhisperX-derived turns) for manual evaluation.[3]

Gold discourse-role labels (TEACHER/STUDENT) were assigned at the turn level based on auditory inspection and interactional function by a single annotator. Because the recordings are dyadic, roles induce a consistent two-speaker mapping per file, standardized as SPEAKER_00 (Student) and SPEAKER_01 (Teacher).

We compare (i) a baseline WhisperX (large-v3) + Pyannote diarization system without role correction and (ii) the full pipeline including LLM-based role correction. Performance is measured as turn-level speaker accuracy. For the full pipeline, predicted and gold turns share identical timestamps and are matched by exact key after rounding to three decimals. For the baseline, gold turns are assigned the speaker label with maximum temporal overlap; non-overlapping cases are excluded.

6.2. Results

The LLM-assisted pipeline substantially improves role attribution over acoustic diarization alone, yielding a 33-point absolute gain in turn-level accuracy on this sampled evaluation set.

[3]We use *turn* to refer to WhisperX segments defined by start/end timestamps and an associated text span.

System	Accuracy (%)
Baseline (WhisperX + Pyannote)	62.70
Full Pipeline (LLM-assisted)	**95.68**

Table 3: Turn-level speaker accuracy on 10 sampled narratives. Baseline labels are mapped by maximum temporal overlap.

7. Implications for Dialect Research and Corpus Construction

The evaluation results suggest that the proposed architecture is not only effective for the MD corpus, but also relevant for other interview-based dialect collections.

7.1. LLM-Assisted Role Attribution in Interview Corpora

Many dialect archives consist of interview-style or field recordings with asymmetric speaker roles.[4] In such settings, acoustic diarization alone often collapses interviewer and respondent speech or inconsistently labels short question turns. Our results suggest that incorporating a lightweight reasoning-oriented LLM step can substantially improve role consistency without manual reannotation.

This approach is directly transferable to other large-scale dialect collections, including 20th-century dialectological or sociolinguistic interviews from the US or UK, where interviewer–informant structure is systematic but acoustically variable.

7.2. Temporal Precision for Variationist Analysis

Dialect research frequently depends on accurate segment boundaries for vowel measurement, duration studies, and turn-taking analysis. By combining WhisperX transcription with MFA-based boundary refinement while preserving original tokenization, the pipeline produces word-level timestamps suitable for corpus-phonetic workflows.

The short alignment-window strategy and token reconstruction mechanism are not specific to Australian English and can be applied to other legacy dialect recordings with comparable acoustic quality.

7.3. Scalability and Cross-Linguistic Potential

The pipeline is modular and language-agnostic in design. While this study focuses on Australian English, the same architecture—ASR, diarization, LLM-based role correction, and forced

[4]For example, field recordings from the 20th-century American *Linguistic Atlas Project*.

alignment—can be adapted to other dialect continua and minority language archives, provided appropriate acoustic and pronunciation models are available.

Because many historical dialect archives remain untranscribed, the approach offers a practical pathway for transforming large volumes of interview data into searchable, temporally aligned corpora. For example, MD_NLP supports geographically indexed extraction of spontaneous-speech tokens in targeted phonological environments, enabling spatial modelling of regional patterning in mid-20th-century AusE using the linked coordinates and informant metadata.

8. Using MD_NLP for Dialect Research

MD_NLP is designed to support workflows that require (i) searchable text, (ii) reliable speaker-role separation, (iii) word-level timing suitable for phonetic extraction, and (iv) linkage to demographic and geographic metadata. Because segments are short and role-consistent, researchers can retrieve targeted stretches of spontaneous speech without extensive manual cleaning.

8.1. Example queries and extraction patterns

The dataset supports common dialectological and corpus-phonetic extraction tasks, including: (i) retrieval of lexical or grammatical patterns conditioned by region (e.g., discourse markers, quotatives, negation patterns), (ii) extraction of phonological environments for vowel/consonant measurement (e.g., pre-lateral contexts, pre-nasal contexts, stress-conditioned subsets), and (iii) modelling of interactional structure (e.g., teacher prompt types vs. student narrative responses).

For example, researchers can (a) filter to STUDENT turns only to avoid interviewer speech, (b) restrict to segments from specific states or subsets of locations using geocoded coordinates, and (c) extract aligned word tokens for downstream forced-alignment or formant measurement pipelines. The availability of school location, birthplace, and parental background variables enables analyses that jointly model regional location and social stratification within the same historical dataset.

8.2. Spatially explicit analyses

Because each segment is linked to geocoded recording sites, MD_NLP can be used in spatially explicit dialectology. Researchers can aggregate measurements (lexical rates, phonetic summaries, discourse-role distributions) at the level of locations, regions, or states and apply spatial statistical methods (e.g., clustering, hotspot analysis, spatial autocorrelation) to test whether observed patterns exhibit geographic structure beyond chance. This is particularly useful for evaluating claims of regional differentiation in AusE, where effects may be subtle and geographically diffuse.

8.3. Diachronic linkage to contemporary corpora

Although MD_NLP focuses on historical narratives, its structure is compatible with modern large-scale speech corpora that provide transcripts and audio. This supports direct comparison of regional patterns across time (e.g., 1959/60 vs. present-day corpora), allowing researchers to test whether regional structure is stable, emerging, or reorganizing. In this sense, MD_NLP can serve as an historical anchor point for contemporary regional studies of AusE.

9. Conclusion

We presented MD_NLP, a discourse-annotated, georeferenced corpus of Mitchell–Delbridge narratives constructed using an integrated ASR, diarization, LLM-based role correction, and forced-alignment pipeline. The resulting resource combines temporally precise word-level timestamps with structured sociolinguistic metadata and national geographic coverage.

Evaluation shows that incorporating a lightweight LLM-based reasoning step substantially improves role attribution in interview-style recordings, increasing turn-level accuracy from 62.70% to 95.68%. This improvement is particularly relevant for dialect corpora in which interviewer–informant asymmetry and variable recording conditions challenge acoustic diarization.

Beyond the MD corpus, the proposed architecture is directly applicable to other large-scale dialect archives, including mid-20th-century North American and British sociolinguistic interviews and comparable legacy collections in other languages. Many such archives remain only partially transcribed and lack reliable speaker attribution; the present approach offers a practical pathway toward transforming them into searchable, temporally aligned, and socially annotated research corpora.

For Australian English specifically, MD_NLP provides new opportunities for investigating regional variability, discourse structure, and corpus-phonetic phenomena in spontaneous mid-20th-century speech at a national scale. By combining alignment precision with rich demographic and geographic metadata, the corpus supports emerging research on spatial and socially conditioned

variation in AusE.

Future work will expand quantitative evaluation, explore additional discourse annotations, and further assess cross-dialect and cross-linguistic transferability of the pipeline.

The integration of structured metadata with discourse-aware speaker attribution positions MD_NLP as a foundation for large-scale, spatially explicit dialect research in Australian English.

Acknowledgements

This work was supported by the European Union – NextGenerationEU instrument and was funded by the Research Council of Finland under grant number 358720. Computational resources were provided by Finland's Centre for Scientific Computing.

10. Bibliographical References

Emily P. Ahn, Gina-Anne Levow, Richard A. Wright, and Eleanor Chodroff. 2023. An Outlier Analysis of Vowel Formants from a Corpus Phonetics Pipeline. In *Interspeech 2023*, pages 2573–2577.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *INTERSPEECH 2023*, pages 4489–4493.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH 2023*, pages 1983–1987.

Luyao Cheng, Hui Wang, Chong Deng, Siqi Zheng, Yafeng Chen, Rongjie Huang, Qinglin Zhang, Qian Chen, Xihao Li, and Wen Wang. 2025. Integrating audio, visual, and semantic information for enhanced multimodal speaker diarization on multi-party conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19914–19928, Vienna, Austria. Association for Computational Linguistics.

Steven Coats, Carmelo Alessandro Basile, Cameron Morin, and Robert Fuchs. 2025a. The YouTube corpus of Singapore English podcasts. *English World-Wide*, 46(3):274–298.

Steven Coats, Chloé Diskin-Holdaway, and Debbie Loakes. 2025b. Regional Distribution of the /el/-/æl/ Merger in Australian English. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 147–156, Abu Dhabi, UAE. Association for Computational Linguistics.

Rolando Coto-Solano, James N. Stanford, and Sravana K. Reddy. 2021. Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems With DARLA. *Frontiers in Artificial Intelligence*, 4.

Felicity Cox and Sallyanne Palethorpe. 2019. Vowel variation in a standard context across four major Australian cities. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, pages 577–581. Australasian Speech Science and Technology Association.

Felicity Cox, Sallyanne Palethorpe, and Samantha Bentink. 2014. Phonetic archaeology and 50 years of change to Australian English /i:/. *Australian Journal of Linguistics*, 34(1):50–75.

Felicity Cox, Joshua Penney, and Sallyanne Palethorpe. 2024. Australian English monophthong change across 50 Years: Static versus dynamic measures. *Languages*, 9(3).

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Jason Li, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Nithin Koluguri, Jocelyn Huang, Jia Fei, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Naren Sean, and Boris Ginsburg. 2023. NeMo: a toolkit for Conversational AI and Large Language Models.

Guillaume Klein and Mahmoud Ashraf. 2023. faster-whisper: Faster whisper transcription with ctranslate2.

Nikola Ljubešić, Peter Rupnik, and Tea Perinčić. 2024. Mići Princ – A Little Boy Teaching Speech Technologies the Chakavian Dialect. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 232–250. Institute of Contemporary History.

Deborah Loakes, John Hajek, and Janet Fletcher. 2017. Can you t[æ]ll I'm from M[æ]lbourne?: An overview of the dress and trap vowels before /l/ as a regional accent marker in Australian English. *English World-Wide. A Journal of Varieties of English*, 38(1):29–49.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger.

2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech 2017*, pages 498–502.

Alexander George Mitchell and Arthur Delbridge. 1998. The speech of Australian adolescents: Research data and recordings collected by AG Mitchell and Arthur Delbridge in 1959 and 1960. The University of Sydney.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH 2023*, pages 3222–3226.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28448–28481. PMLR.

Penelope Schmidt, Chloé Diskin-Holdaway, and Debbie Loakes. 2021. New insights into /el/-/æl/ merging in Australian English. *Australian Journal of Linguistics*, 41(1):66–95. _eprint: https://doi.org/10.1080/07268602.2021.1905607.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. 2024. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models. In *Interspeech 2024*, pages 3754–3758.

Challenges in the Detection of Dialect for Historical Languages; the Case of Old Irish Text Resources

Adrian Doyle
Maynooth University
Department of Early Irish
School of Celtic Studies
adrian.odubhghaill@mu.ie

Abstract

Old Irish presents particular challenges for the study of automatic dialect detection. It is generally accepted that little trace of dialect can be found in Old Irish writings. Extant Old Irish text resources introduce a considerable amount of extra variation, which could impact dialect identification applications. While some scholarship has suggested that certain features may be indicative of dialect, such hypotheses are difficult to substantiate where authorship is anonymous, or where the text itself is not associated with a particular geographical region. This paper describes the application of stylometric dialect detection techniques to Old Irish texts, and discusses the features which emerge from this process as potential markers of dialect. The aim is not necessarily to identify Old Irish dialectal features with certainty, but rather to investigate the impact that Old Irish text resources could have on such applications. This paper does, nevertheless, add to the extant body of research by highlighting features which have been identified as stylistically distinct by stylometric techniques intended for dialect detection.

Keywords: Old Irish, Early Irish, Historical Languages, Dialect Detection

1. Introduction

For most of the 20th century, conventional wisdom in the field has been that Old Irish, the Gaelic language which survives from roughly the 7th to the 10th centuries, does not show significant signs of dialectal variation. Forms of variation certainly do occur in texts surviving from this period, and in 1985, McCone noted that "peculiarities" and "deviations" in the Old Irish glosses had "been known about for a long time". Such forms of variation as exist, however, have generally been attributed to reasons other than dialect. Thurneysen writes, for example, that "Linguistic differences in the Old Irish sources are almost all differences of period, and are the result of morphological development" (1946, 12). In other words, the language can be seen developing between the earliest and latest writings from the period, however, "Contemporary divergences, such as would point to dialectal peculiarities, are very rare..." (1946, 12). While affirming this position more recently, stating "The language encountered in Old Irish texts shows a surprisingly high degree of uniformity, with hardly any dialectal distinctions discernible," Stifter (2006, 10) goes on to assert that, nevertheless, "these certainly must have existed at the time." This echoes O'Rahilly's earlier claim that "it is not rash to assume that the Irish spoken in those times by the common people was not quite uniform in every part of the country" (1932, 248). Though O'Rahilly too admits that, "Of dialect differences in Old and Middle Irish we know nothing or next to nothing", earlier in the same work he makes the comparatively radical, and apparently contradictory statement, that "A perusal of the following pages will lead to the conclusion that historically there were but two main dialects in Irish, a Northern and a Southern..." (1932, 16).

Aside from the relatively small amount of Old Irish text surviving in contemporary sources, the problem of detecting Old Irish dialect is compounded by forms of text variation introduced at the editorial level. The orthographic representation of the language as it appears in manuscripts has been shown to be difficult to digitise accurately (Doyle et al., 2018), while the lack of a single digital editorial standard (Dereza et al., 2023), and of a single tokenisation or word separation standard (Doyle and McCrae, 2025a; Doyle et al., 2019), have been demonstrated to impact natural language processing (NLP) applications for Old Irish in other task areas. As has been discussed in Doyle and McCrae (2025a; 2025b), several extant repositories of Old Irish text follow their own discrete editorial and word separation guidelines (Griffith, 2013; Lash, 2014; Bauer et al., 2023; Stifter et al., 2021). This introduces variation between resources which is not indicative of original authorship in any way, and hence could likely impact stylometric analyses applied to texts drawn from more than a single text repository. Meanwhile, the quantity of text, and variety of sources, which have been edited in adherence to a single methodology (Doyle, 2018, 2023a,b) are too limited to be useful in a dialect detection experiment.

Section 2 of this paper will discuss the back-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 33–47
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ground to the problem of dialect in the field of Early Irish studies in more detail, while section 4 will discuss applications of stylometric techniques for the detection of dialect in other language areas. Section 3 will discuss existing text resources for Old Irish and their suitability to such stylometric applications. Finally, section 5 will describe an experiment carried out whereby stylometric analysis for the detection of dialects was applied to a selection of Old Irish texts drawn from the repository found to be most suitable for this type of application. The aim is not only to outline what the application of such techniques might suggest about authorship in the case of Old Irish, but to make the case that Old Irish text resources should be constructed in such a manner that they better facilitate the application of techniques such as these.

2. Background

It has been suggested that writings surviving in Old Irish may reflect a literary standard (Stifter, 2006, 10) rather than any particular dialect. Stifter (2009, 60) expands on this notion:

> There is little or no trace of synchronic variation in the Old Irish literary tradition, what variation exists being mostly stylistic rather than geographical (Kelly, 1982; McCone, 1989). This presupposes either the early adoption of a specific local variety as the basis for a standard, or the early codification of a standard grammar. The sporadic appearance already in the glosses of features of phonology, morphology and syntax which only become prominent in the Middle Irish period after the tenth century (McCone, 1985), suggests that the dominant register in these texts is a conservative literary standard at some remove from the spoken language, and perhaps one generation older than the earliest attested texts.

This description of Old Irish writings outlines two obstacles to the detection of dialect in the language. Firstly, little synchronic variation can be found. Secondly, even where synchronic variation may be apparent, it may not be indicative of dialect. Such variation could, for example, reflect a distinction in linguistic register, or scribal variation whereby a scribe copying an Old Irish text into a later manuscript may deliberately or inadvertently corrupt, alter or modernise some forms. Scribes may even introduce hypercorrections, attempting to reproduce unfamiliar, archaic forms, but introducing aberrations in the process. In order to be sure that any form of synchronic variation is indicative of dialect, therefore, it is necessary to be able to assign a geographical location to the variant in question. Murray (2005, 100) notes that this problem "is exacerbated by the fact that much of the material under investigation is anonymous."

Despite claims that "The language encountered in Old Irish texts shows a surprisingly high degree of uniformity..." (Stifter, 2006, 10), it should be noted that a significant amount of variation does nevertheless occur in Old Irish texts. Spelling variation, for example, is not uncommon in writings in historical languages, and the use of diacritics to mark long vowels and lenition in Old Irish specifically is inconsistent. *Corpus Palaeohibernicum* (CorPH) (Stifter et al., 2021), identifies 539 types of variation which occur within that text repository alone, and each of these is grouped into one of five distinct variation categories; phonological, orthographical, morphological, syntactical, and lexical[1]. The extent to which variation of any of these kinds might be useful in the identification of dialect requires further study, however, a handful of specific types of variation have been outlined in the literature as potential dialectal markers.

Kelly (1982) notes that, beside terms which are regularly used for certain animals, rarer words can also be found to refer to the same animals. Kelly posits that this is unlikely to be diachronic variation, as common terms, such as *bó* "ox/cow", *ech* "horse", *mucc* "pig", *cáera* "sheep", *gabor* "goat", and *sinnach* "fox", are used in the earliest Old Irish texts, and so cannot simply be later replacements for earlier terms which still appear, albeit rarely (1982, 86). Ahlqvist (1988, 34) later suggested 'it may be important to realise that the importance of the "prestigious forms of Old Irish" is largely a function of modern scholars' knowledge about the language that was codified in Thurneysen's (1946) Grammar', underscoring the potentially important role that non-standard forms, such as those outlined by Kelly, might play in identifying markers of Old Irish dialect. Accordingly, Ahlqvist points to variation in the form of the anaphoric pronoun (specifically *ón* versus *són*) in the Old Irish glosses as another potential marker of dialect.

An argument which may reasonably be levelled against either of these hypotheses is that, even if these forms of variation do represent genuine dialectal differences, it is nevertheless impossible to be sure of this, unless any of these forms can be reliably linked to a particular geographical region. In response to this, Murray (2005, 106) suggests that a study of "Placenames would seem to offer the most hope for geographically locating possible dialectal features in medieval Irish vocabulary."

[1] https://chronhib.maynoothuniversity.ie/chronhibWebsite/tables/variations?page=0&limit=100&fprop=&fval=&dtable=variations&ctable=&search=false

He argues that not only does variation in placenames allow for geographical localisation, as variants can be accurately pinpointed, but also that "placenames are liable to fossilise early forms of the language", thereby preserving quite old variants, without much fear they may have been standardised or modernised.

In one of the most comprehensive studies in this topic area to date, Malthaner (2022) claims to have achieved two aims. Firstly, "to seek out and examine synchronic variation within the Old Irish record" (2022, 273), and secondly, to identify "the existence of dialects within the material". Malthaner claims to have identified "a significant amount of variation", and hence, claims to have "dismantled the long-held academic belief that the Old Irish period was free from linguistic variation." Though she concedes that "... while a number of features were identified and investigated, the majority were found to not be conclusive evidence of diatopic variation" (2022, 274), Malthaner nevertheless suggests that "at least one feature [has] been determined to be a diatopic variant" (2022, 273). Malthaner also dismisses "the theory of Common Gaelic (Jackson, 1951) which has pervaded scholastic thought ... owing to the unlikelihood of an entirely invariant language being able to exist for such a significant period of time" and argues for the position that Old Irish was "a scholastic standard that was utilised by an educated elite ... deviations from which can be understood to be slips into more natural registers".

A significant amount of orthographic and script variation can occur in Old Irish manuscripts, even among the writings of individual scribes. When such writings are digitised, differing editorial standards can result in further variation being introduced, and several ways in which extant digital corpora for Old Irish are distinct from one another as a result of these factors are discussed in Doyle and McCrae (2025a, 1–4). While automatic dialect identification has been demonstrated to be effective in other languages (Alsuwaylimi, 2024; Bernier-Colborne et al., 2022; Imaizumi et al., 2022), the text resources available for Old Irish limit the types of techniques which are applicable, and it is conceivable that irrelevant forms of variation could impact applications which are otherwise feasible. As such, the following questions are worth asking:

1. How might the lack of editorial standardisation in Old Irish text resources impact such applications?

2. How might orthographic and script variation in manuscripts impact such applications?

3. What might be identified as potential markers of dialect by applicable NLP techniques?

4. How will any potential dialectal features which are automatically identified compare with extant research on Old Irish dialects?

3. Resources

Before question 1, posed in the preceding section, can be addressed, it is necessary to briefly discuss the types of text resources which are available for use in experiments like the one discussed in this paper. This will give an indication of the type and quantity of text data which is available from existing resources, as well as the state of non-standardisation between them. As has been noted by Dereza et al. (2023), no text editing standard exists for digital Old Irish editions. This results in editorial variation between texts which can impact downstream NLP applications. It has also been noted that word separation, and hence tokenisation, are not carried out in a uniform manner across different text resources for Old Irish either (Doyle and McCrae, 2025a). While it is not feasible to discuss the extent of this non-standardisation here, relevant examples will be provided below to demonstrate how this can be problematic.

The majority of text surviving in contemporary sources from the Old Irish period comes in the form of glosses. These typically take the form of short interlinear notes in manuscripts. Three large corpora exist comprising thousands of glosses, these are the Würzburg, Milan and St. Gall gloss collections. Otherwise, a small amount of prose and poetry also survives. Each of these three collections of glosses have been digitised (Griffith, 2013; Doyle, 2018; Bauer et al., 2023). Another resource, CorPH (Stifter et al., 2021), is the largest repository of annotated Early Irish texts, and incorporates two of these collections, Milan and St. Gall, along with many other Early Irish texts. All of these resources provide morphological annotation for their contents[2], however, no two of these resources make use of the same editorial standard or annotation style. Similarly, the first text repository containing Old Irish to be parsed, POMIC (Lash, 2014), uses a discrete POS tag-set (Doyle and McCrae, 2025a, 2). Finally, two Universal Dependencies (UD) treebanks exist, each containing a small number of glosses (Doyle, 2023a,b). As is discussed in Doyle and McCrae (2025b), each of these make use of the same editorial standard and POS tag-set as the digitised Würzburg collection (Doyle, 2018), however, with fewer than 100 parse trees each, these are two of the smallest treebanks in the UD collection. While all of the resources mentioned here provide some level of

[2]Though annotation is not complete for the Würzburg collection, it is stated in Doyle and McCrae (2025b, 395) that over 600 glosses have been annotated to date.

lexical annotation, it will be unsurprising given the current state of research into Old Irish dialects, that no existing corpus is annotated with labels to indicate dialectal differences.

While data annotation standards enforced by UD, alongside the diplomatic editing methodology employed during their creation (Doyle and McCrae, 2025b), might suggest that the Old Irish UD treebanks would be good candidates for use in an experiment like the one presented here, unfortunately, there simply isn't enough annotated data available in these resources. Similarly, while POMIC does contain texts from various sources, and there is somewhat more data than is available in UD treebanks, the total quantity is still quite small. Thus, though some research has suggested that POS-tags and syntactic data can have a predictive power in stylometric analyses (Eder and Górski, 2022), the only parsed corpora for Old Irish do not contain enough data to support meaningful analysis in this case.

Looking at the remaining annotated text data which is available for Old Irish, it becomes clear that any data used must be drawn from a single resource. Where different editorial standards are employed, it is shown in Doyle and McCrae (2025a, 2) that this can result in variation even between the same text content as presented by different resources. It is also demonstrated in Doyle and McCrae (2025a) that, even where raw text is initially the same between resources, when word separation is applied this can result in drastically different tokens being produced by different resources. This type of variation might easily be expected to impact stylometric analyses applied to such resources. While there may be a sufficient quantity of text data in any one of the three individual gloss collections (Griffith, 2013; Doyle, 2018; Bauer et al., 2023) to conduct stylometric analysis, if all text data is drawn from a single manuscript source the chances of finding meaningful, potentially dialectal distinctions would be significantly reduced. Thus, the best option is to utilise data from CorPH, the largest of the annotated resources, comprising texts drawn from multiple sources, all annotated using a single methodology.

While it is not possible to link the language used in any of these texts with any certainty to any particular geographical area, many of the manuscripts themselves have been linked to particular regions, or scriptoria. For example, the annals of Ulster and writings from the Book of Armagh, can reasonably be linked geographically with the general Ulster area. Similarly Hofman (1996, 19–23) argues that the manuscript containing the St. Gall glosses was created in either Nendrum or Bangor, hence, this can also be linked with Ulster. While manuscripts such as these may indeed represent a local variety of the language, this kind of geographical association alone does not necessarily ensure this, as scribes writing in these locations may have come to the area from elsewhere. Nevertheless, selecting texts linked with a specific region for use in the experiment described here remains the best available option for potentially isolating dialect markers.

4. Related Work

A variety of NLP approaches have demonstrated the potential for identifying dialect across a variety of languages. Alsuwaylimi (2024) discusses the use of two hybrid bidirectional long short-term memory models for dialect detection in Arabic, Bernier-Colborne et al. (2022) describe the use of multi-class support vector machine classifiers, as well as a probabilistic classifier for French, and Imaizumi et al. (2022) discuss a multi-task learning approach for Japanese dialect identification and multi-dialect automatic speech recognition using a transformer-based system. Several factors separate these approaches from the case of Old Irish, not least the quantity and standardisation of text data which is typically available for modern languages. The chief distinguishing factor, however, is that for each of these languages, it is already possible to identify specific dialects, and link them to geographical regions. As such, models can be trained and tested on dialect-labelled data. This is not possible for Old Irish, and so another approach is necessitated.

Several publications suggest that stylometric analyses can be employed reliably to identify features like dialect (Lahjouji-Seppälä et al., 2022), and register (Grieve, 2023). Lahjouji-Seppälä et al. (2022) demonstrate the utility of a stylometric approach to analyse the diachronic development of 20th century Standard Ukrainian, and Grieve (2023) suggests that stylometric analysis is consistent with standard theories of register variation. Stylometric methods are even demonstrated to be useful in separating AI generated Arabic text from that of humans, including dialectal social media posts, by identifying stylistic features consistent with large language model text generation (Al-Shaibani and Ahmed, 2025). Moreover, Goswami et al. (2020) describe a method for simultaneously learning sentence embeddings and clustering assignments from short texts for the purpose of dialect identification, advancing the state-of-the art in both supervised and unsupervised settings. This is particularly relevant for texts like Old Irish glosses, which are typically very short.

A stylometric approach engineered to separate texts based on dialectal features is the best available approach for an Old Irish application, as it does not necessitate the use of dialect-labelled

data for training or testing of models. While this means that the accuracy of such an approach cannot be tested for Old Irish, that is not the primary aim. Instead, the main focus is to determine what impact variation within extant text resources will have on the application of an experiment intended to detect dialect. A stylometric approach will allow the text resource to be treated as a variable, with the same approach being applied to raw text, and to word-separated text from the same repository. As has been discussed in Doyle and McCrae (2025a), in Old Irish orthography spacing is not inserted systematically between all words, and the application of word separation in CorPH results in the addition of many characters, and even duplication of many morphs[3], such that the resulting tokens cannot be directly aligned with the original text. As such, it is possible to apply a single stylometric technique to both the original text, and to the recombined tokens, to demonstrate the impact such editorial variation can have on such an application. This, in turn, can indicate how the standard of data available from text resources could impact other dialect detection applications for Old Irish.

5. Experiment

To date, traditional scholarship has been focussed primarily on the lexical level, however, dialectal difference can manifest in orthography, morphology, syntax and sub-word character sequences. Even looking at lexical choice, however, use of function words, particles and clitics can be telling, but these do not receive as much focus in the literature for Old Irish. The experiment described below applies a stylometric clustering approach intended to identify latent dialect structure in a collection of short texts, without relying on labelled training data or predefined dialect categories. This approach should be capable of identifying systematic differences in lexical choice, orthography, and morphosyntactic patterns which, among other things, may be indicative of dialect.

5.1. Methodology

Taking each sentence, gloss, or verse in the corpus to be a text instance, the aim is to find statistical regularities across these instances. Each instance is assumed to be capable of encoding diagnostic dialectal cues (such as variant spellings, function words, particles, clitic forms) alongside other, irrelevant forms of variation. Texts were drawn from CorPH only, and all text instances were extracted from only the following list of texts:

1. Annals of Ulster
2. Armagh Additamenta
3. Armagh Glosses Minor Glosses
4. Armagh Liber Angeli
5. Armargh Notulae [sic]
6. Milan Donatus Minor Glosses
7. Milan Glosses
8. Milan Priscian
9. Milan Sententia Sanctorum Doctorum Minor Glosses
10. Poems of Blathmac
11. St Gall Priscian
12. St Gall Prudentius Minor Glosses

These specific texts were chosen either because they can be somewhat reliably ascribed to a named geographical location, can be ascribed to a named author (hence, they represent a single dialect), or because they are one of the large collections of Old Irish glosses. All text instances were added to a list, and labelled with the name of the text from which they were taken. These labels were not used during clustering, and were retained for visualisation purposes only.

Sentence embedding was carried out using a pre-trained, multilingual sentence embedding model[4] from the `SentenceTransformers` framework. This pre-trained model was chosen as no labelled dialect data was available. Having been pre-trained on large multilingual corpora, the aim was to avail of the model's transfer learning capabilities. This should allow the model to apply its general understanding of semantic and syntactic structure to the task of encoding Old Irish text instances, even without having ever been trained on Old Irish. The use of sentence embeddings allows for syntactic structure to be taken into account as well as lexical features. Text instances are mapped to 768-dimensional vector space by this model, and cosine distance is used to compute similarity between instance vectors.

[3] CorPH refers to elements produced during word-separation in that corpus as "morphs" rather than "tokens". However, as many of these "morphs" represent entire words, including any inflections, this use of the term is atypical. The term "morph", where used in this paper, is intended in the more traditional sense of a word segment which represents the smallest unit of language that has meaning. To avoid confusion, the "morphs" of CorPH will be referred to here as tokens.

[4] `paraphrase-multilingual-mpnet-base-v2` https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Unsupervised clustering was carried out using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN has a few benefits which make it more useful for this application than other clustering methods. Firstly, because the number of potential dialects is unknown, it was not desirable to predefine a number of clusters. Secondly, HDBSCAN labels text instances which do not clearly belong to any dense region as noise rather than assigning them to any cluster. With potentially thousands of noisy text instances, this allows for irrelevant instances to be removed so that only relevant clusters are represented when the points are plotted. The minimum cluster size was set to 15 and minimum samples were set to 5 to ensure a lower number of clusters, increasing the likelihood that clusters will represent something closer to the scale of dialect, rather than that of authorship or turns of phrase.

Dimensionality reduction was carried out using UMAP (McInnes et al., 2020). UMAP was chosen for its ability to preserve both local and global structures better than alternatives like t-SNE, while compressing noise and redundant dimensions. Moreover, because dialect continua may be less likely to form linearly separable clusters than other stylometric classifications, UMAP might be expected to better capture gradual transitions between varieties as it approximates the underlying manifold structure of the data. Embeddings are projected to two dimensions after dimensionality reduction, and plotted. Each point plotted represents an individual text instance which the HDBSCAN process deemed significant enough to be assigned to a specific cluster, i.e. not noise. Plotted instances are colour coded to identify them with the corpus they were originally drawn from.

5.2. Variants

Four further variants of this experiment were also carried out to see how altering certain variables would affect results. Each of these variants were set up as per the methodology outlined above, with the only difference being to the variable in question.

5.2.1. Variant 1: TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) was used instead of a sentence embedding model to quantify the importance of individual terms in a document given the frequency with which they appear in that document. The TF-IDF for a given term, *t*, in a given document, *d*, taken from a collection of many documents, *D*, is calculated as follows:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

As the primary focus of this approach is on the lexical level, measuring the frequency with which specific strings of characters (i.e. terms) occur within a document, it may be more sensitive to spelling variation than an embedding model. For a historical language like Old Irish, the considerable amount of arbitrary variation typical of texts might therefore impact the usefulness of a TF-IDF approach in an application like this. On the other hand, the ability of embedding models to detect semantically similarity, and hence interpret two texts as saying the same thing even when written differently, may cause them to overlook genuinely meaningful forms of variation in a task such as this.

The minimum document frequency for a term was set to 2, and the n-gram range was set between 1 and 2 to allow the measure to take syntactic context into account also. Using the `TfidfVectorizer` from `scikit-learn` (Pedregosa et al., 2011), TF-IDF embeddings were produced. All other parameters remained the same as in the methodology described above.

5.2.2. Variant 2: Character-level TF-IDF

Because word-level TF-IDF only looks at the frequency with which word-tokens appear, it cannot account particularly well for features like spelling variation, morphology and punctuation, all of which might indicate dialectal variation. For this reason, character level TF-IDF was also applied. This looks at combinations of characters rather than entire words within a text instance. The minimum document frequency was again set to 2, but the n-gram range was set between 2 and 5. Otherwise parameters were as per the first variant, above.

5.2.3. Variant 3: t-SNE

Because some amount of separation seemed apparent between clusters for certain corpora, while other clusters appeared crowded together (see discussion below), one variant of this experiment was carried out using t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction instead of UMAP. According to van der Maaten and Hinton (2008), t-SNE "produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map." The hope was that this method might separate some of the more crowded clusters apart in the plot, making it easier to visualise potential points of dialectal difference. Perplexity was set to 30, and cosine similarity was used to measure high-dimensional similarities between data points. Otherwise, the embedding model, and all other variables, were as per the methodology described above.

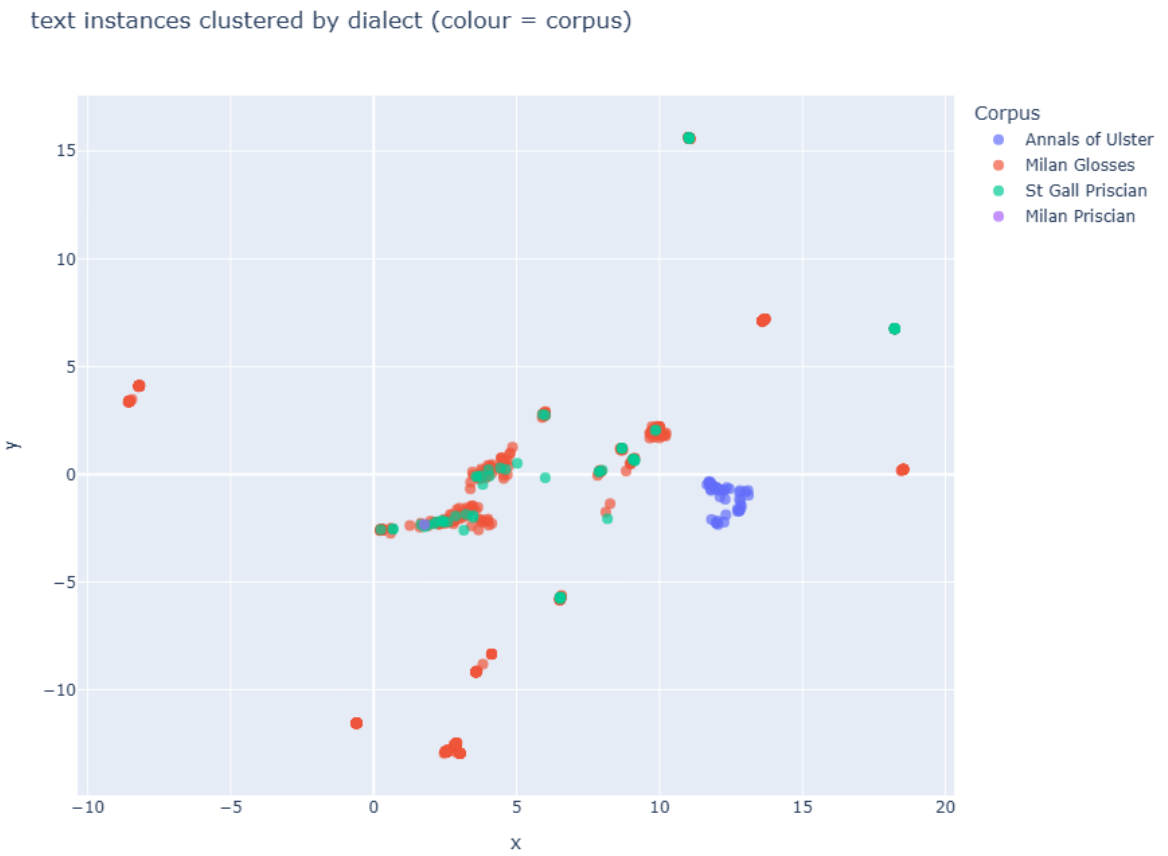


Figure 1: Plot of Dialect Clusters for Old Irish Using Sentence Embeddings.

5.2.4. Variant 4: Recombined Tokenised Text

As has been discussed in section 3, tokens resulting from CorPH's word-separation process do not necessarily match the original text exactly if recombined. A fourth variant of the experiment was set up to investigate how this kind of preprocessing would affect results. CorPH tokens for each text instance were concatenated with spaces between each token. While this may not result in new text instances which resemble actual Old Irish text very closely, words are nevertheless clearly separated, which might make it easier for the embedding model to identify differences at the lexical level which could indicate dialect. To exemplify the distinction, one text instance drawn from the Milan Glosses (Ml. 113a9), and used in all other experiment variants, reads *duróscai* in its original form. Recombining the tokens resulting from CorPH's word-separation as described here, however, results in the string, "`du· ·ró ós du·róscai`". Aside from the text input, all other variables and parameters for this experiment variant were as described in the methodology section above.

5.3. Results and Discussion

Figure 1 shows the results of the clustering process described in the methodology section above, after text instances were plotted. What is immediately clear is that a lack of data impacted the experiment. Once noisy data was removed, clusters were found to be comprised of data from only four collections; Annals of Ulster, Milan Glosses, St Gall Priscian, and Milan Priscian. As can be seen, contents from the Annals of Ulster form their own cluster to the centre-right of the plot, while another large cluster to the centre-left is comprised of data from the St. Gall Priscian, the Milan Glosses, and a single data point from the Milan Priscian. It must be kept in mind that these clusters could indicate stylistic features other than dialect, for example, linguistic register or text subject matter. Whatever the underlying factors though, it is significant that in a stylometric experiment geared towards dialect identification, these clusters remain after noise removal, and they are the largest two clusters. They account for more than simple similarity in lexical choice.

Several smaller clusters can be found around the edges of the plot, comprised of text instances from the Milan Glosses and St Gall Priscian. These are very telling of the impact that both manuscript and editorial variation can have on an experiment of this sort. Clusters 10 (-1, -12), 12 (-8.5, 4.5), 13 (3, -12.5) and 15 (1, -2.5) are primarily comprised of variant forms of the same word, with spelling variation. All instances in cluster 10 are either *.i. adǽ* or *ostú .i. adǽ*, and all instances in 12 are, *adæ*. Many instances of *.i. adæ* and *.i. adae* were lumped into the leftmost end of cluster 15, though the cluster itself continues as far to the right as (4, -1.5), and contains a wide variety of other text instances also. Finally, cluster 13 comprises the forms *adǽ, ádǽ, adé,* and *ádæ*. The fact that these forms are numerous enough to comprise four clusters, suggests that this term may be a significant feature of style in the Milan Glosses, however, the fact that they do not all appear in a single cluster suggests that irrelevant variation has impacted these results. It is possible that standardisation of manuscript spelling variation could result in clearer findings.

None of the features of style plotted in these clusters can be clearly likened to anything in the dialectal literature for Old Irish to date. Nevertheless, the considerable overlap between the St. Gall Priscian and Milan Glosses would seem to confirm arguments that little dialectal variation can be found between these Old Irish sources. While both are corpora of glosses, their subject matter is very different. St. Gall deals with the grammar of Latin, while Milan relates to the Psalms. As such, the overlap cannot be attributed simply to similarity in genre, and there does appear to be some stylistic similarity shared between these two collections which is not shared by the Annals of Ulster. While the separation of the Annals of Ulster from the gloss corpora in the plot might be significant stylistically, it may equally reflect the inherent distinction in style between glosses and annal entries. Alternatively, it could also reflect stylistic differences indicating diachronic language variation rather than dialectal.

Because of limited space, plots displaying the results for the experiment variants can be found in the Appendix.

5.3.1. Results for Variant 1: TF-IDF

The results from the first variant of the experiment can be seen in Appendix A. In this variant, TF-IDF was used to measure stylistic similarity rather than an embedding model. Comparing these results against those in Figure 1, this method appears to have had more difficulty separating text instances into meaningful clusters. The result is a single, large grouping of data points towards the right of the plot, containing text instances from all plotted corpora, including the Annals of Ulster. It may be noteworthy, however, that after noise reduction text instances from more corpora remained in this plot than that of the embedding model. As with Figure 1, the same outlying, small clusters can be found here, i.e. *adǽ*, *.i. adǽ*, *adæ*, etc.

5.3.2. Results for Variant 2: Character-level TF-IDF

Appendix B shows the results of the second variant of the experiment, which utilised character-level TF-IDF as a measure. Here, the Annals of Ulster are separated relatively clearly again into their own cluster, centred around (1, 1). Again, instances from most other corpora form into a relatively large grouping towards the bottom, centre of the plot, with small, outlying clusters surrounding. One notable feature of this plot is that content from the Poems of Blathmac is relatively tightly grouped around the (6.3, 8.6) mark, seemingly indicating either strong markers for the poet's own personal style, or simply a distinction in genre between poetry, glosses, and annal entries.

5.3.3. Results for Variant 3: t-SNE

As predicted, clusters are spread further apart in the plot resulting from dimensionality reduction using t-SNE (Figure 4). This results in several smaller clusters comprised primarily of content from both St. Gall Priscian and the Milan Glosses. Outliers from the other plots now form much tighter clusters at (50, -35), (85.5, 6.5), and (93,-13.5). Otherwise, the contents of the Annals of Ulster remain separated from those of other corpora, comprising a single cluster at the centre, left. This plot does make the visualisation of distinct clusters more intuitive, though the overlap of two or more corpora in several large clusters underscores the fact that the underlying embedding model still considers these corpora to be stylistically similar.

5.3.4. Results for Variant 4: Recombined Tokenised Text

The extra lexical information available to the model in the fourth variant of the experiment appears to have resulted in much tighter groupings, with clusters being more definitively separated from each other in the plot (Figure 5). Still, there remains significant crossover between the contents of the St. Gall Priscian and Milan Glosses corpora, while content from the Annals of Ulster again forms an entirely separate cluster in the bottom centre of the plot. Even the outliers remain the same, albeit, with spacing now introduced between elements. Cluster 22 (17.8, 1.5) is comprised of *a dǽ, a dé,* and *.i. a dé*, 11 (23.5, 6) is comprised of *a dæ* and

á dæ, while 12 (21.3, 2) is entirely comprised of *.i. a dæ*.

6. Future Work

A major limitation of the work presented here is that the accuracy of any of the models described cannot be quantified, as this would require existing Old Irish text resources to be already annotated with dialect labels. In lieu of this, a valuable avenue for future research would be to apply the same stylometric techniques described in this paper to the modern Gaelic languages. Because dialectal distinctions in Modern Irish are well understood, for example, it would be possible to assess the accuracy of stylometric techniques on such a corpus. Though modern and historical forms of Irish are distinct in many ways, and so it is not possible to extrapolate that any measured accuracy would hold between the two, this would nevertheless allow researchers a better understanding of the relationship between clusters and genuine dialects in a very closely related language. This knowledge could then be used to improve understanding of applications to Old Irish, informing the tuning of hyperparameters and reading of results.

Another potentially valuable area for future investigation into dialect detection for Old Irish would be training of models to identify texts known to have been produced in different geographical areas. This would allow the accuracy of the dialect detection tool to be assessed in a way which was not possible here. If sufficiently accurate, features deemed to be likely representative of dialect by such a model could then be examined. To create such a model, however, would first require the production of digital editions of texts like the West Munster Synod, which are known to have been produced in areas which are geographically distinct from those which were available for use in this research. More importantly, as this research has shown, it would be important that the production of such texts is standardised to minimise the impact of editorial variation on downstream NLP techniques like dialect detection.

7. Conclusion

This paper has conducted a stylometric dialect identification experiment on Old Irish text. The aim was not to identify specific features which can be certainly assigned to dialect in Old Irish, as separation of clusters alone cannot be taken as an indication of dialect. Rather, the aim here was to find features of style which do separate corpora, so that these features can be examined in more detail by experts to determine the reasons clusters were separated. Where corpora like the Annals of Ulster, or the Poems of Blathmac – which can be linked to either a geographical region or an individual author – form tight clusters in the plots, this may indicate the presence of some stylistic features indicative of dialect. This, however, will require further examination to determine. Nevertheless, the results presented here appear to show a clear distinction, in several experiment variants, between a single cluster containing text instances from the Annals of Ulster, and several other clusters in which text instances from the Milan Glosses and St Gall Priscian are regularly mixed. Unfortunately, too little data was available for other corpora for their contents to have been equally heavily represented in any of the plots after noise was removed, though some methods examined here clearly resulted in more aggressive noise removal than others.

Another pressing aim of this research was to determine the impact of editorial non-standardisation on the application of techniques such as this to a historical language, like Old Irish. It was clear that plots differed even where the same techniques were applied, and the only variable was the standard of the text. Word-separated text, which was recombined using spacing, produced clusters which were much more clearly defined than those produced using the original text. This underlines the impact that editorial decisions can have. A relatively major finding was that, even in the case of an apparently significant stylistic feature, editorial non-standardisation resulted in the generation of up to four spurious text clusters where, at most, one was required. As such, this paper reiterates the point made by Dereza et al. (2023), that it is important to to investigate the potential value of deciding a digital text editing standard for Old Irish to increase the efficacy of NLP applications to the language.

To summarise the value of the work presented here, the primary achievement has been in showing that the only currently viable techniques which might be applied to Old Irish texts in an effort to identify dialectal features can be heavily impacted by the editorial standards applied in the production of text resources. The paper has also shown a strong tendency for text instances from two large gloss corpora to be clustered together, while annal entries remained separated throughout most of the experimentation conducted. Similarly, the existence of several small clusters unique to only one of the large gloss collections suggests some stylistic features unique to that corpus, whether they be dialectal or not. All of these preliminary findings may be useful in informing future dialectal investigations by more traditional means, by drawing attention to stylistic features which have not been the focus of attention to date.

8. Acknowledgements

I would like to express my gratitude to Prof. David Stifter for kindly reviewing this paper prior to publication, and for providing very helpful insights.

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number IRCLA/2023/2124.

9. Limitations

As mentioned above, not enough data was available, edited and annotated to the same standard. The experiment would have benefitted from the inclusion of the third large corpus of Old Irish glosses, Würzburg, however, these are not yet available in CorPH. Other corpora which can be assigned to specific locations outside of Ulster would have also improved the outcome of this experiment, for example, if the West Munster Synod were available through CorPH.

10. Bibliographical References

Anders Ahlqvist. 1988. Remarks on the Question of Dialects in Old Irish. In Jacek Fisiak, editor, *Historical Dialectology: Regional and Social*, pages 23–38. De Gruyter Mouton, Berlin, New York.

Maged S. Al-Shaibani and Moataz Ahmed. 2025. The Arabic AI Fingerprint: Stylometric Analysis and Detection of Large Language Models Text.

Amjad A. Alsuwaylimi. 2024. Arabic Dialect Identification in Social Media: A Hybrid Model with Transformer Models and BiLSTM. *Heliyon*, 10(17).

Gabriel Bernier-Colborne, Serge Leger, and Cyril Goutte. 2022. Transfer Learning Improves French Cross-Domain Dialect Identification: NRC @ VarDial 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 109–118, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Oksana Dereza, Theodorus Fransen, and John P. McCrae. 2023. Do not Trust the Experts - How the Lack of Standard Complicates NLP for Historical Irish. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 82–87, Dubrovnik, Croatia. Association for Computational Linguistics.

Adrian Doyle and John P. McCrae. 2025a. An Assessment of Word Separation Practices in Old Irish Text Resources and a Universal Method for Tokenising Old Irish Text. In *Proceedings of the 5th Celtic Language Technology Workshop*, pages 1–11, Abu Dhabi [Virtual Workshop]. International Committee on Computational Linguistics.

Adrian Doyle and John P. McCrae. 2025b. Development of Old Irish Lexical Resources, and Two Universal Dependencies Treebanks for Diplomatically Edited Old Irish Text. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 393–402, Albuquerque, USA. Association for Computational Linguistics.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 2018. Preservation of Original Orthography in the Construction of an Old Irish Corpus. In *Proceedings of the LREC 2018 Workshop: "CCURL2018 – Sustaining Knowledge Diversity in the Digital Age"*, pages 67–70, Miyazaki, Japan.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 2019. A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles. In *Proceedings of the Celtic Language Technology Workshop*, pages 70–79, Dublin, Ireland. European Association for Machine Translation.

Maciej Eder and Rafał L. Górski. 2022. Stylistic Fingerprints, POS-tags, and Inflected Languages: A Case Study in Polish. *Journal of Quantitative Linguistics*, 30(1):86–103.

Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. Unsupervised Deep Language and Dialect Identification for Short Texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1606–1617, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jack Grieve. 2023. Register Variation Explains Stylometric Authorship Analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77.

Rijcklof Hofman. 1996. *The Sankt Gall Priscian Commentary. Part 1. Volume 1: Introduction; Book 1-5*. Nodus, Münster.

Ryo Imaizumi, Ryo Masumura, Sayaka Shiota, and Hitoshi Kiya. 2022. End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning. *APSIPA*

Transactions on Signal and Information Processing, 11(1):1–23.

Kenneth Jackson. 1951. 'Common Gaelic': the Evolution of the Goidelic Languages. *Proceedings of the British Academy*, 37:71–97.

Patricia Kelly. 1982. Dialekte im Altirischen? In W. Meid, H. Ölberg, and H. Schmeja, editors, *Sprachwissenschaft in Innsbruck*, pages 85—89. Innsbrucker Beiträge zur Kulturwissenschaft, Innsbruck.

M.Z. Lahjouji-Seppälä, A. Rabus, and R. von Waldenfels. 2022. Ukrainian Standard Variants in the 20th Century: Stylometry to the Rescue. *Russian Linguistics*, 46(3):217–232.

Ariana Nicole Malthaner. 2022. *Synchronic Language Variation in the Old Irish Glosses*. Trinity College Dublin, Dublin. PhD Thesis.

Kim McCone. 1985. The Würzburg and Milan Glosses: Our Earliest Sources of 'Middle Irish'. *Ériu*, 36:85–106.

Kim McCone. 1989. Zur Frage der Register im frühen Irischen. In Stephen N. Tranter and Hildegard L. Tristram, editors, *Mündlichkeit und Schriftlichkeit in der frühen irischen Literatur*, pages 57–97. Gunter Narr Verlag, Tübingen.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Kevin Murray. 2005. Dialect in Medieval Irish? Evidence from Placenames. *Studia Celtica Fennica*, 2:97–109.

Thomas F. O'Rahilly. 1932. *Irish Dialects Past and Present*. Dublin Institute for Advanced Studies, Dublin.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

David Stifter. 2006. *Sengoidelc*. Syracuse University Press, New York.

David Stifter. 2009. Early Irish. In Martin J. Ball and Nicole Müller, editors, *The Celtic Languages*, chapter 4, pages 55–116. Routledge, Abingdon, New York.

Rudolf Thurneysen. 1946. *A Grammar of Old Irish*, 2 edition. The Dublin Institute for Advanced Studies, Dublin.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

11. Language Resource References

Bauer, Bernhard and Hofman, Rijcklof and Moran, Pádraic. 2023. *St Gall Priscian Glosses, version 2.1*.

Doyle, Adrian. 2018. *Würzburg Irish Glosses*.

Doyle, Adrian. 2023a. *Diplomatic St. Gall Glosses Treebank*. Universal Dependencies.

Doyle, Adrian. 2023b. *Diplomatic Würzburg Glosses Treebank*. Universal Dependencies.

Griffith, Aaron. 2013. *A Dictionary of the Old-Irish Glosses*.

Lash, Elliott. 2014. *The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1*. The Dublin Institute for Advanced Studies.

Stifter, David and Bauer, Bernhard and Lash, Elliott and Qiu, Fangzhe and White, Nora and Barrett, Siobhán and Griffith, Aaron and Bulatovas, Romanas and Felici, Francesco and Ganly, Ellen and Nguyen, Truc Ha and Nooij, Lars. 2021. *Corpus PalaeoHibernicum (CorPH) v1.0*. National University of Ireland, Maynooth.

Appendix

A. Experiment Variant 1: TF-IDF

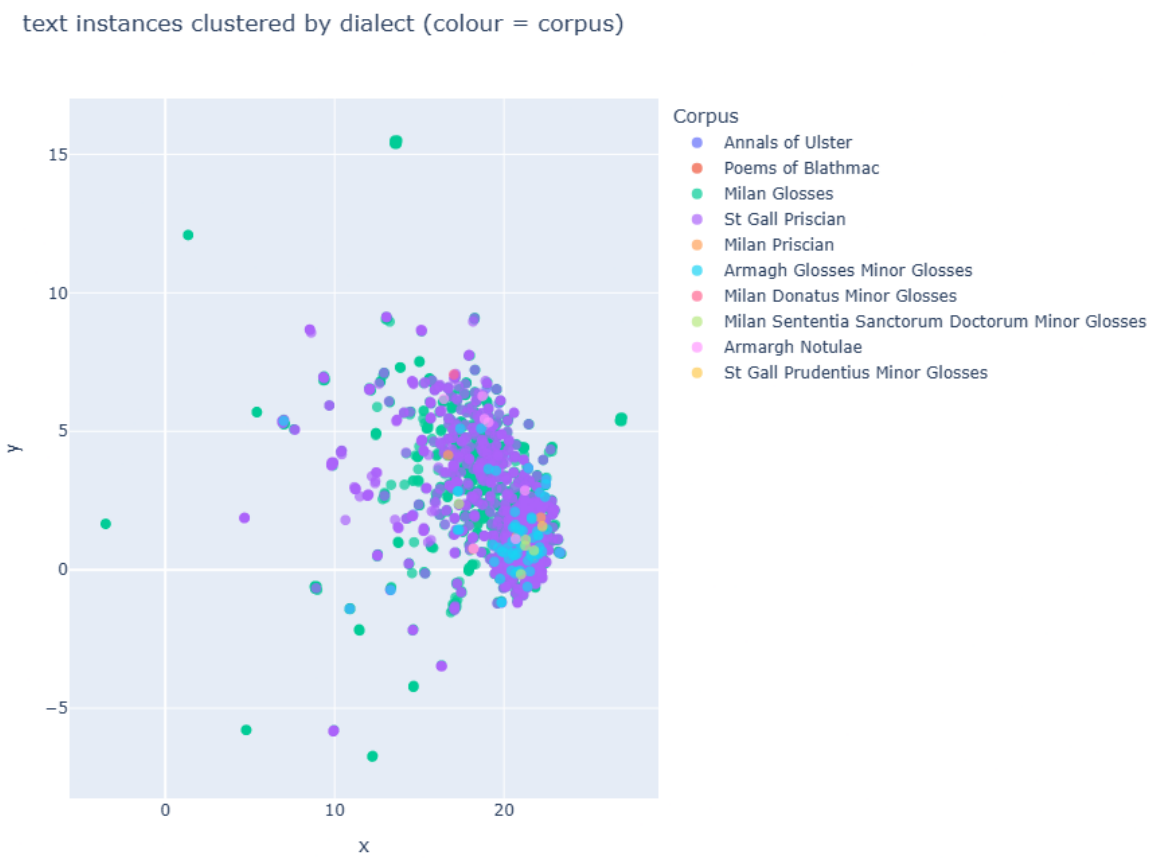


Figure 2: Plot of Dialect Clusters for Old Irish using TF-IDF.

B. Experiment Variant 2: Character-level TF-IDF

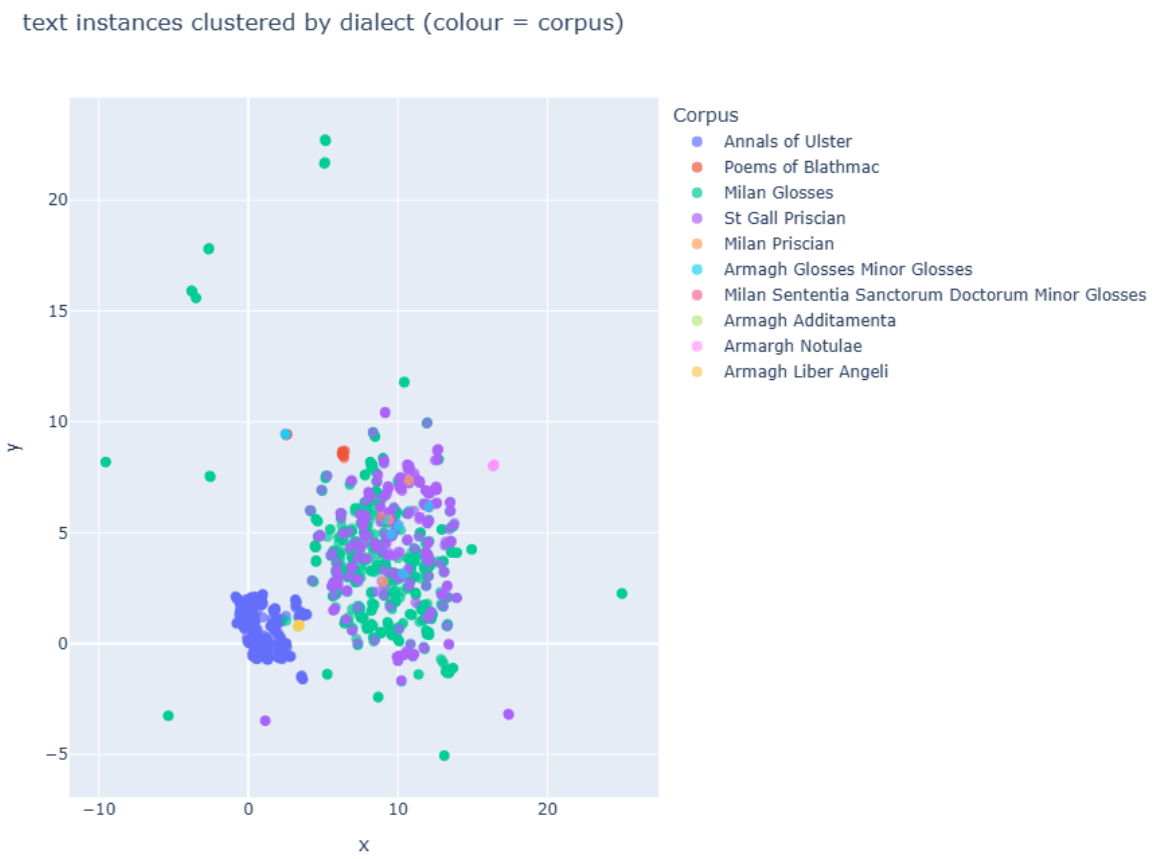


Figure 3: Plot of Dialect Clusters for Old Irish using Character-level TF-IDF.

C. Experiment Variant 3: t-SNE



Figure 4: Plot of Dialect Clusters for Old Irish using t-SNE.

D. Experiment Variant 4: Recombined Tokenised Text

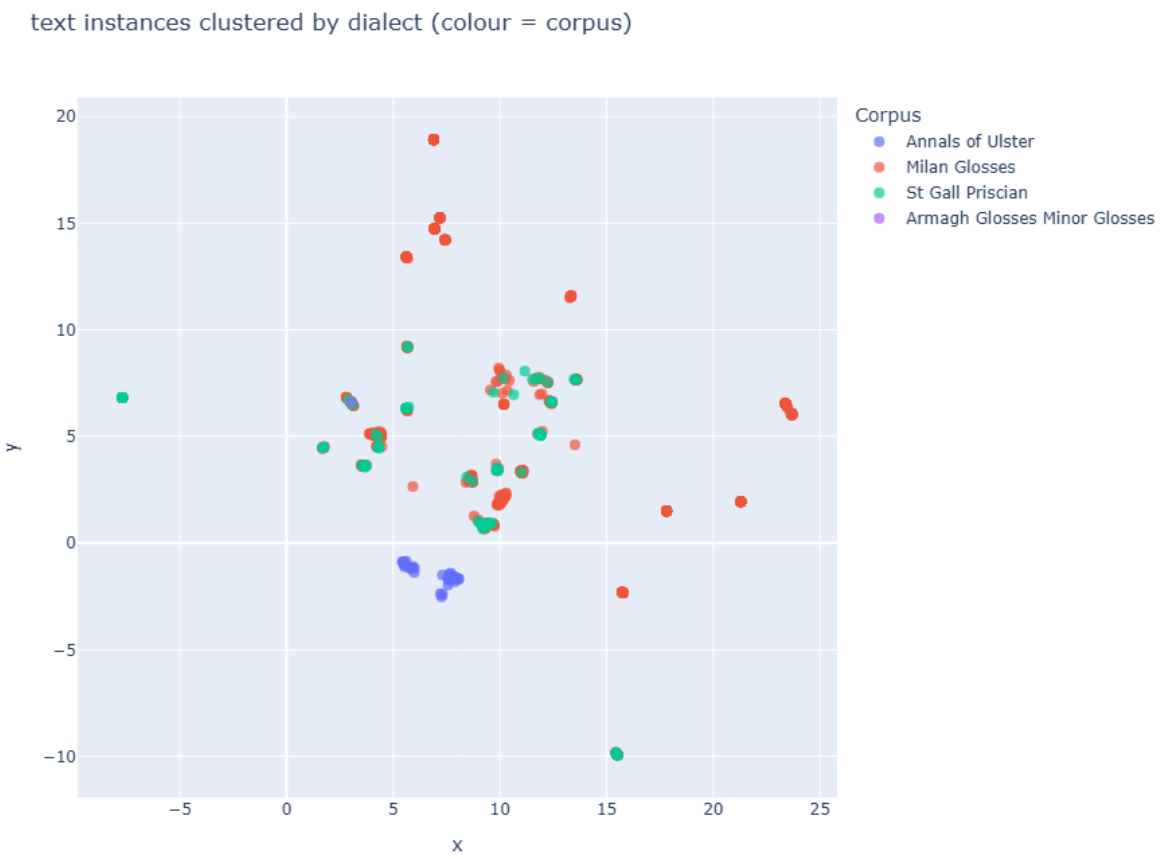


Figure 5: Plot of Dialect Clusters for Word-separated Old Irish Text.

Phonologically-aware Automatic Speech Recognition Evaluation of Low-Resource Languages: The Case of Basque Dialects

Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas, Inma Hernáez
HiTZ Basque Center for Language Technology - Aholab Signal Processing Laboratory,
University of the Basque Country UPV/EHU, Spain
{christoforos.souganidis, asier.herranz, ibon.saratxaga, eva.navas, inma.hernaez}@ehu.eus

Abstract

Automatic speech recognition models are typically trained with data of standard languages. However, their performance degrades when dealing with non-standard dialectal speech. In this paper, we present the first evaluation of an automatic speech recognition system for Basque, a low-resource language, based on spontaneous broadcast speech with high representation of dialectal speech. It relies on a 140-h manually annotated proprietary corpus of television programs in Basque, including dialect-level labels, as well as standardized and pseudo-phonetic transcriptions. We find that recognition performance significantly degrades for dialectal compared to standard speech, for all dialects present in our corpus. Subsequently, we provide a quantitative analysis of phonological phenomena based on single-word substitution errors, and identify 52 recurrent phenomena, grouped into sound deletions, epentheses, and substitutions. We further show a modest but statistically significant correlation between the number of phonological phenomena in an utterance and its recognition error rate. Our findings highlight the limitations of dialect-agnostic evaluation and motivate linguistically informed, dialect-aware strategies for automatic speech recognition in low-resource and typologically diverse languages.

Keywords: automatic speech recognition, dialectal speech evaluation, Basque dialects

1. Introduction

The vast majority of language technologies have been trained with standardized varieties of languages. However, dialectal variation has gained increasing attention in natural language processing (NLP), as existing systems frequently struggle to handle socially conditioned linguistic variation such as dialects, underscoring the need to systematically integrate social factors into NLP models and applications (Hovy and Yang, 2021; Zampieri et al., 2020). However, a recent survey on dialects in NLP (Joshi et al., 2025) emphasizes that despite the social and ethical implications of dialectal NLP, most of the applications focus on high-resource languages with many speakers, such as English or Spanish (Zampieri et al., 2024). A great recent advance towards the incorporation of language varieties in NLP was the creation of DIALECTBENCH, the first large-scale benchmark for dialectal NLP (Faisal et al., 2024), and the dialectal evaluation of machine translation benchmark CODET (Alam et al., 2024).

In addition to affecting NLP, social factors, such as race or dialectal variation, are found to affect the performance of automatic speech recognition (ASR) models. The impact of accent on ASR performance has troubled the field for over two decades (e.g. Huang et al., 2004), with phonology playing a crucial role (Yang et al., 2018). Various studies show that, up to date, ASR performance degrades for speech of non-white speakers or speakers of non-standard varieties (Koenecke et al., 2020; Tatman, 2017; Wassink et al., 2022; Chang et al., 2024; Vakirtzian et al., 2025; Martin and Tang, 2020). As in the case of NLP, ASR advances in dialectal speech focus mainly on English (Hinsvark et al., 2021), leaving other languages underexplored, although recent benchmarking initiatives, namely ML-SUPERB 2.0 (Chen et al., 2025; Shi et al., 2024), have started to integrate dialectal speech.

1.1. Basque dialectal landscape

Basque is a language isolate spoken in the Autonomous Community of the Basque Country and Navarre in Spain, and in the Northern Basque Country in France. Although it has approximately only 1 million speakers (Basque Government, 2016), there is a great deal of dialectal variability in terms of phonetics and phonology (Hualde, 1991; Hualde and Ortiz de Urbina, 2003). Various proposals have been made regarding the classification of Basque dialects, ranging from 3 to 8, many of which contain subdialectal divisions (Zuazo, 2008; Aurrekoetxea et al., 2017, 2019). In this paper, we adopt a dialectal classification based on the division of Basque-speaking regions into provinces: Biscayan (biz), Gipuzkoan (gip), Upper Navarrese (naf), Lapurdian (lap), Lower Navarrese (naf-low), and Souletin. The Royal Academy of the Basque language delevoped Standard Basque (eus), the standardized version of the Basque language, based on central dialects, mainly on Gipuzkoan, incorporating features of

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 48–57
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

other dialects such as Lapurdian and Upper Navarrese (Hualde and Zuazo, 2007). This is also reflected in recent NLP studies, showing lower Levenshtein distance between Standard Basque and dialects closer to Standard Basque (i.e., Gipuzkoan and Upper Navarrese), than between the standard and more peripheral dialects such as Biscayan (Bengoetxea et al., 2025).

Regarding phonetics and phonology in Basque dialects, we can classify the phenomena into three major categories: deletions, epentheses and substitutions. Deletions of sounds can occur in various word positions, such as between vowels (intervocalic deletions), for example, <egun> ('day') > <eun>, or in the final position of a syllable (coda deletions), as in the case of <usain> ('smell') > <usai> (Hualde et al., 2010). Consonant (C) epentheses are typically attested in an intervocalic position, for example, <mendia> ('mountain') > <mendixa> (Hualde et al., 2010). Substitutions can vary in form, such as substitution of a vowel (V) with another vowel or vowels, substitution of consonants including palatalization, and voicing assimilation. Basque dialects show a wide variety of vocalic interactions, for example <neska> ('girl') that can be pronounced either as <neske>, <neskia> or <neskie> (Hualde and Gaminde, 1998; deCastro Arrazola et al., 2015). Some Basque dialects maintain the voiceless sibilant distinction between apico-alveolar <s> and laminal <z>, while others merge the two sounds (Hualde and Gaminde, 1998; Larraza et al., 2017). Palatalization in Basque orthography is denoted by the use of a double consonant, typically <tt> or <dd> (palatalized versions of <t> and <d>, respectively), as in the case of <polita> ('pretty') > <politta> (Hualde et al., 2010). Voicing assimilation of voiceless consonants is common when the preceding sound is voiced, for example, in the expression <lan eta lan> ('working non-stop') which is pronounced <lan da lan>, after the common deletion of <e> in <eta> ('and') in spoken language.

As illustrated by the phonological phenomena described above, the linguistic landscape of Basque dialects is highly complex. This is particularly relevant for computational resources of dialects of an already low-resource language, as the lack of specific subdialectal orthography makes it difficult for the representation of those dialects in NLP and speech technologies (e.g. Vakirtzian et al., 2025).

1.2. Basque dialectal resources

Despite Basque being considered a low-resource language, recent efforts are starting to include dialectal variation, reflecting the growing interest in Basque NLP and speech technologies.

Regarding NLP resources of Basque dialects, there have been a handful of contributions: a historical corpus of Basque dialects (Estarrona et al., 2020), a corpus of syntactic variation (Uria and Etxepare, 2012), the XNLI-var dataset, including Biscayan, Gipuzkoan, and Upper Navarrese dialects (Bengoetxea et al., 2025), and the recent BasPhyCo$_{west}$ for commonsense reasoning in Biscayan (Bengoetxea et al., 2026). Although not specifically targeted as a resource for dialectal NLP, the corpus of tweets by Fernandez de Landa (2019) includes dialectal material from Biscayan and Gipuzkoan varieties, which was later integrated as part of the Diverse Corpora corpus as an NLP resource aimed at capturing language variation (Azurmendi et al., 2025).

Compared to the increasing availability of dialectal resources for NLP for Basque, dialectal speech processing resources remain relatively underexplored, with a couple of notable exceptions. The Speech_Dat (II) telephone speech database (Hernaez et al., 2003) contains untranscribed recordings of Standard and Biscayan Basque speakers from 23 regions. The Basque Speecon-like and Basque SpeechDat MDB-600 databases (Odriozola et al., 2014) contain manually transcribed spontaneous and read speech by 832 speakers from the Autonomous Community of the Basque Country and Navarre in Spain. Nevertheless, in these datasets there is an under-representation of spontaneous speech, and, as in the case of Speech_Dat (II), include only Basque speakers from Spain. In a study comparing performance of ASR in Basque and Spanish, it is found that performance for Basque is noticeably worse than for Spanish, and although it is not systematically investigated, dialectal variability is likely a contributing factor for this discrepancy (Peñagarikano et al., 2023).

1.3. Aims of this work

Despite recent advances in Basque dialect resources for NLP, there is still a lack of studies looking at the performance of ASR systems for Basque dialects. In particular, we hypothesize that the variable phonological phenomena attested in Basque dialects could be responsible for the behavior of ASR systems.

In this study, our contributions are twofold:

- a dialect-aware evaluation of Basque ASR of spontaneous speech on real broadcast data, and
- a systematic documentation of phonological phenomena in various Basque dialects.

This work provides empirical evidence on ASR performance across Basque dialects, and documents relevant phonological phenomena in spoken Basque, contributing to both dialect-aware ASR research and linguistic studies of dialectal variation.

Dialect	$N_{speakers}$	$N_{utterances}$	N_{words}	H
eus	924	63565	489620	59
lap	27	812	7905	1
gip	547	49103	296365	33
biz	335	29791	170043	19
naf	42	2073	15290	2
naf-low	1	8	28	<1

Table 1: Number of speakers, number of utterances, number of words, and number of hours, for each dialect.

2. Data

In all experiments carried out in this paper, we used a proprietary dataset of TV shows in Basque broadcast by Basque Radio Television (Euskal Irrati Telebista, EITB), the Autonomous Community of the Basque Country's public broadcast service. The dataset consists of 140 hours of manually annotated and transcribed audio from speakers of various Basque dialects.

It contains two types of transcription: a standardized transcription and a pseudo-phonetic transcription. The standardized transcription corresponds to the written form of each utterance, either in Standard or Dialect-Specific Basque. However, due to the aforementioned dialectal diversity even within each dialect, the limits of what constitutes a standardized transcription are sometimes difficult to define. The pseudo-phonetic transcription corresponds to the orthographic representation of what was pronounced by each speaker. This type of transcription includes phonological phenomena such as across-word or within-word co-articulation.

The audio data was annotated for speaker turns, speaker identity (each speaker was assigned a unique label), and speaker dialect (one dialect label per speaker). Table 1 shows the number of speakers per dialect. It should be noted that the dialect annotation denoted the dialect identified for each speaker, regardless of the content of each utterance. This means that even if an utterance is pronounced by a speaker of a dialect and a speaker of Standard Basque in exactly the same way, the former is annotated as dialectal, while the latter is annotated as standard. The speakers whose dialect could not be identified were annotated as speakers of Standard Basque. This corpus contains the following dialects: Biscayan, Gipuzkoan, Upper Navarrese, Lapurdian and Lower Navarrese.

3. Analyses and results

The statistical analyses described in this section were carried out in JASP v0.19.1.0 (JASP Team, 2026). A non-parametric Kruskal–Wallis test was conducted, with the evaluation metric as the dependent variable and *Dialect* as the independent variable. Since the independent variable had more than two levels, we conducted Dunn's post-hoc pairwise comparisons using Holm correction of *p*-values.

3.1. Evaluation of ASR

For the experiments conducted in this study, we used a Conformer-Transducer ASR model developed within the NVIDIA NeMo[1] framework. The specific model used was `stt_eu_conformer_transducer_large_v2`, which is publicly available on the Hugging Face Hub[2]. The standardized transcriptions were used as reference for the ASR evaluation.

The model was trained on a total of 772 hours of Basque speech. The training corpus includes a diverse collection of speech data, with 95.75 hours of spontaneous speech sourced from broadcast radio programs, distinct from the programs in our proprietary dataset. Notably, this training set incorporates specific dialectal representation, including 19 hours of Biscayan, 11.7 hours of Gipuzkoan, 3.45 hours of Lapurdian and 1 hour of Upper Navarrese dialects. This inclusion of dialect-rich broadcast data allows the model to better generalize across the phonological variations characteristic of the Basque linguistic landscape described in Section 1.1.

The descriptive statistics (Table 2) show that the ASR system had better performance in terms of word error rate (WER) for standard compared to dialectal speech. The non-parametric Kruskal-Wallis test with *WER* as the dependent variable and *Dialect* as the independent variable showed a significant effect of *Dialect* on *WER* (*H(5)* = 6630.784, $p < .001$). Pairwise comparisons (Table 3) revealed that performance for Standard Basque was significantly better than for all other dialects ($p = .005$ for Lower Navarrese, $p < .001$ for the rest of the dialects). The only dialect that was significantly more challenging for ASR compared to all other dialects except Lower Navarrese was Gipuzkoan. It should be noted that although Lower Navarrese had a numerically higher WER, the limited amount of observations as shown in Table 2 was not enough for a statistically reliable difference.

3.1.1. Substitution patterns

In this section, we focus on the single-word substitutions produced by the ASR model described above, as our goal is to explore the effect of dialectal speech on ASR performance. Since we are aiming to investigate the phonological phenomena that

[1]NeMo: https://github.com/NVIDIA/NeMo

[2]Model: https://huggingface.co/HiTZ/stt_eu_conformer_transducer_large_v2

Dialect	$\text{WER}_{\text{utt}} \pm \text{SD}$	$\text{WER}_{\text{global}}$
eus	0.192 ± 0.324	0.126
lap	0.245 ± 0.283	0.213
gip	0.350 ± 0.407	0.270
biz	0.332 ± 0.406	0.256
naf	0.318 ± 0.413	0.227
naf-low	0.619 ± 0.301	0.643

Table 2: Mean WER and standard deviation per utterance, and global WER, per dialect.

Comparison	z	p_{holm}
eus − lap	−9.406	< 0.001
eus − gip	−76.136	< 0.001
eus − biz	−54.375	< 0.001
eus − naf	−16.544	< 0.001
eus − naf-low	−3.422	0.005
lap − gip	−3.525	0.004
lap − biz	−1.382	0.501
lap − naf	−0.887	0.750
lap − naf-low	−2.496	0.095
gip − biz	10.302	< 0.001
gip − naf	3.930	< 0.001
gip − naf-low	−2.128	0.133
biz − naf	0.549	0.750
biz − naf-low	−2.342	0.106
naf − naf-low	−2.373	0.106

Table 3: Dunn's post-hoc pairwise comparisons for *Dialect*, reporting *z* statistics and Holm-corrected p-values.

might affect ASR, the focus of the following evaluation shifts to the character error rate metric (CER). We believe that CER reflects more accurately than WER the direct effect of dialectal variation on ASR performance, as WER could be influenced by other factors such as the number of dialect-specific words per utterance.

To specifically analyze substitution patterns across dialects, we focused the analysis on single-word substitution errors. First, we filtered out all the correctly recognized utterances by our ASR model, as well as the word deletions and word insertions. Then, for the remaining data, we kept only single-word substitutions that differed between the manual transcriptions and the automatic transcriptions and calculated the CER, considering the manual transcriptions as the ground truth.

Table 4 displays the number of single-word substitutions, as well as the mean CER and standard deviation per dialect of these substitutions. In this table, we can observe that the CER in Standard Basque is the lowest compared to dialectal speech. The non-parametric Kruskal-Wallis test with *CER* as the dependent variable and *Dialect* as the independent variable showed a significant effect of *Dialect* on *CER* (*H(5)* = 652.503, $p < .001$). As

Dialect	N	$\text{M} \pm \text{SD}$
eus	22993	0.344 ± 0.300
lap	498	0.361 ± 0.281
gip	26293	0.385 ± 0.330
biz	14020	0.394 ± 0.321
naf	1033	0.378 ± 0.341
naf-low	6	0.372 ± 0.338

Table 4: Number of utterances, mean CER and standard deviation of single-word substitutions per dialect.

Comparison	z	p_{holm}
eus − lap	−2.799	0.056
eus − gip	−21.293	< 0.001
eus − biz	−22.158	< 0.001
eus − naf	−4.862	< 0.001
eus − naf-low	−0.084	1.000
lap − gip	−1.448	1.000
lap − biz	−2.427	0.137
lap − naf	−0.511	1.000
lap − naf-low	0.225	1.000
gip − biz	−4.319	< 0.001
gip − naf	1.186	1.000
gip − naf-low	0.387	1.000
biz − naf	2.568	0.102
biz − naf-low	0.497	1.000
naf − naf-low	0.294	1.000

Table 5: Dunn's post-hoc pairwise comparisons for *Dialect*, reporting *z* statistics and Holm-corrected *p*-values.

shown in Table 5, CER was significantly lower in Standard Basque compared to most dialects ($p < .001$), namely Gipuzkoan, Biscayan and Upper Navarrese, which were the dialects with the highest amount of data in our corpus. Between dialects, the only significant difference found was between gip and biz, with the latter showing a significantly higher CER ($p < .001$). Our results are aligned with the distance of each dialect from the standard, with Upper Navarrese being closer, followed by Gipuzkoan, and Biscayan being the most different (Bengoetxea et al., 2025).

3.2. Phonological analysis

In order to investigate the phonological phenomena that could affect ASR performance, we compared the standardized and the pseudo-phonetic transcriptions available in this corpus. First, we identified the single words in the corpus that had a different standardized transcription compared to the pseudo-phonetic one. Then, we kept only unique transcription pairs and calculated the frequency of occurrence of each pair. We identified 53541 pairs, 13412 of which were unique. We manually anno-

tated the phonological phenomena of 367 unique pairs that occurred more than 15 times. These manual annotations correspond to a total of 31343 occurrences, approximately 58,54% of all pair instances.

We identified 52 phenomena (see Table 7), which can be grouped into three: deletions ($n_{deletions} = 6$), epentheses ($n_{epentheses} = 6$), and substitutions ($n_{substitutions} = 40$).

Deletions. We identified two types of deletion, namely consonant (C) deletions and vowel (V) deletions. We further classified C-deletions depending on the position of the deleted sound, either in syllabic coda (without sound specification) or in intervocalic position (including sound specification). For example, if the word ‹ezagutu› ("to get to know"), was pronounced ‹ezautu›, the pair was annotated with the phenomenon *intervocalic g deletion*.

Epentheses. We identified two types of epentheses, i.e., single C and syllabic. Single C epentheses included the specification of the sound, while syllabic ones did not. For example, if the word ‹seiehun› ("six hundred"), was pronounced ‹seirehun›, the pair was annotated with the phenomenon *intervocalic r epenthesis*.

Substitutions. Three types of substitutions were identified: CV-to-CV, C-to-C, V-to-V. For C-to-C and V-to-V substitutions, the C/V could be either a single C/V or two C/V. In cases where only one C/V was involved in the substitution but was adjacent to another C/V which did not change, both C or V letters were annotated as part of the phenomenon. As an example, the word ‹orduan› ("then") pronounced as ‹orduen›, the pair was annotated as *ua-to-ue substitution*. It should be noted that, despite the fact that we annotated the phenomena involving two C/V as substitutions, these could also be viewed as cases of deletions or epentheses. Regarding C-to-C substitutions, two of the phenomena identified were palatalization and voicing of a voiceless stop sound. Although some C-to-C substitutions could also be viewed as palatalizations (e.g., *s-to-x substitution*), we annotated as *palatalization* only cases where a C is written as a double C, which in Basque are palatalized.

In some pairs, more than one phenomenon was identified, which were either considered independent or dependent. In the latter, we considered that some phonological phenomena were applied recursively, particularly where a deletion of a sound could have triggered a subsequent substitution. For instance, if the word ‹lehenago› ("earlier") was pronounced ‹lehenau›, the pair was annotated with the phenomenon *intervocalic g deletion*, which triggered the phenomenon *ao-to-au substitution*.

We identified 51 phenomena in Gipuzkoan, 50 in Biscayan, 31 in Upper Navarrese, 15 in Lapurdian, while Lower Navarrese was not present in our annotation due to insufficient data. Of all the identified phenomena, only one was unique to a single dialect: *oa-to-ue substitution* was only present in Biscayan. Finally, we explored the co-occurrence of phenomena that were unique to a specific dialect: The co-occurrence of *e-to-i substitution* and *intervocalic t deletion* was unique to Biscayan, while *intervocalic g deletion* and *intervocalic r deletion* co-occurrence and the co-occurrence of *intervocalic g deletion*, *ea-to-ia substitution*, and *ei-to-i substitution* were unique to Gipuzkoan.

Dialect	M ± SD
lap	1.000 ± 0.000
gip	1.161 ± 0.395
biz	1.362 ± 0.498
naf	1.200 ± 0.404

Table 6: Mean number of phenomena per word (M) and standard deviation (SD), per dialect.

Last, we explored whether the number of attested phenomena differs between dialects. Table 6 displays the descriptive statistics. A non-parametric Kruskal-Wallis test with the Number of phenomena as the dependent variable and Dialect as the independent variable showed a significant effect of Dialect (H(3)= 101.029, p < .001). As Dialect had four levels, we conducted Dunn's post-hoc pairwise comparisons using Holm correction of p-values. Biscayan showed significantly more phenomena compared to all other dialects, i.e., Gipuzkoan (z = 9.829 , p < .001), Upper Navarrese (z = 2.603, p = .037), and Lapurdian (z = 3.685, p = .001).

3.3. Linking phonological variations and ASR errors

In this section, we bring together the results from the ASR evaluation regarding Basque dialects and our phonological analyses. We investigated if ASR performance in terms of CER is mediated by the number of identified phenomena in utterances with single word substitutions. For this reason, we conducted a Spearman's correlation analysis between the total CER and the number of phenomena in the utterance. Our results suggest a modest but significant correlation (r = .216, p < .001).

4. Discussion

This study provides, to our knowledge, the first dialect-aware analysis of Basque ASR performance on spontaneous broadcast speech. The results consistently show that recognition errors are not uniformly distributed across varieties, as Standard Basque shows significantly better performance in terms of WER and CER compared to

Group	Phenomenon	Dialects attested	Example		
			Standardized transcription	Pseudophonetic transcription	Dialect
deletion	coda	biz, gip, naf	bat	ba	biz
deletion	e	biz, gip, lap, naf	honetan	hontan	gip
deletion	intervocalic g	biz, gip, lap, naf	egin	ein	biz
deletion	intervocalic r	biz, gip, lap, naf	gero	geo	gip
deletion	intervocalic t	biz, gip, naf	dute	due	gip
deletion	intervocalic d	biz, gip, lap, naf	didazu	diazu	gip
epenthesis	d	biz, gip	genuen	gendun	gip
epenthesis	n	biz, gip, naf	hola	holan	biz
epenthesis	r	biz, gip, naf	seiehun	seirehun	biz
epenthesis	syllable	gip, naf	izan	izandu	naf
epenthesis	x	biz, gip	egia	egixa	biz
epenthesis	z	biz, gip	dira	diez	biz
substitution	a-to-ai	biz, gip, lap	zergatik	zergaitik	biz
substitution	a-to-au	biz, gip, lap, naf	handia	haundia	gip
substitution	a-to-e	biz, gip, lap, naf	behala	bezela	gip
substitution	a-to-o	biz, gip	daukat	dakot	biz
substitution	ao-to-au	biz, gip	dago	dau	gip
substitution	ao-to-o	biz, gip, lap, naf	lasaiago	lasaio	gip
substitution	au-to-a	biz, gip, lap, naf	daukazue	dakazue	gip
substitution	da-to-te	biz, gip	dodala	dotela	biz
substitution	e-to-a	biz, gip, lap, naf	hemezortzi	hamazortzi	biz
substitution	e-to-i	biz, gip, naf	behar	bihar	gip
substitution	ea-to-e	biz, gip	behar	ber	gip
substitution	ea-to-i	biz, gip	azkenean	azkenin	gip
substitution	ea-to-ia	biz, gip, lap, naf	batean	batian	gip
substitution	ea-to-ie	biz, gip, naf	azkenean	azkenien	biz
substitution	ei-to-e	biz, gip, naf	hogeita	hogeta	gip
substitution	ei-to-i	biz, gip, naf	begira	bira	gip
substitution	eo-to-o	biz, gip	edo	o	gip
substitution	ia-to-i	biz, gip	guztian	guztin	gip
substitution	ia-to-ie	biz, gip	egia	egie	gip
substitution	i-to-e	biz, gip, lap, naf	nirekin	nerekin	gip
substitution	ie-to-i	gip, naf	horiek	hoik	gip
substitution	kl-to-k	biz, gip	klaro	karo	gip
substitution	n-to-l	biz, gip	lehenengo	lelengo	gip
substitution	o-to-u	biz, gip, naf	non	nun	gip
substitution	oa-to-o	biz, gip	orain	oin	biz
substitution	oa-to-u	biz, gip, naf	joatea	jutea	gip
substitution	oa-to-ua	biz, gip, naf	goazen	guazen	gip
substitution	oa-to-ue	biz	gaurkoan	gaurkuen	biz
substitution	oe-to-o	biz, gip, lap, naf	dagoen	dagon	biz
substitution	ou-to-u	biz, gip	moduz	muz	gip
substitution	palatalization	biz, gip, naf	ditu	dittu	gip
substitution	s-to-x	biz, gip	samarra	xamarra	biz
substitution	t voicing	biz, gip, naf	eta	da	gip
substitution	u-to-eu	biz, gip	nuke	neuke	gip
substitution	u-to-o	biz, gip, naf	justu	justo	biz
substitution	ua-to-u	biz, gip, lap, naf	orduan	ordun	gip
substitution	ua-to-ue	biz, gip, naf	orduan	orduen	biz
substitution	ue-to-u	biz, gip, lap, naf	genituen	genitun	gip
substitution	z-to-s	biz, gip	ezta	esta	biz
substitution	z-to-x	biz, gip, naf	dezente	dexente	gip

Table 7: Examples of dialectal phonological phenomena with the dialect of the example and the dialects where each phenomenon is attested.

most dialectal varieties. This finding aligns with earlier observations in other languages, where ASR systems trained predominantly on standardized varieties struggle to generalize to socially and regionally marked speech (Koenecke et al., 2020; Tatman, 2017; Wassink et al., 2022; Chang et al., 2024; Martin and Tang, 2020; Vakirtzian et al., 2025).

A particularly relevant outcome is the systematic disadvantage observed for Biscayan and Gipuzkoan, which are both well represented in our dataset, being the two dialects with the highest number of utterances. The fact that Biscayan exhibits significantly higher CER than Gipuzkoan, despite high data availability, suggests that the observed differences cannot be explained by data volume alone. Instead, our phonological analysis supports the interpretation that the linguistic distance to the standard variety plays a central role. This interpretation is further reinforced by the finding that Biscayan utterances display a significantly higher number of phonological phenomena per word than the other dialects considered, consistent with previous research (Bengoetxea et al., 2025).

The significant correlation between the number of phonological phenomena in an utterance and its CER provides evidence that phonological variability contributes to recognition errors. Although the correlation is not strong, this is expected given that ASR errors are influenced by multiple interacting factors, including background noise and speaker characteristics. Nevertheless, our results indicate that the accumulation of phonological processes is a relevant predictor of ASR performance in dialectal Basque.

Importantly, our qualitative analysis reveals that most of the attested phenomena are not dialect-exclusive. Only one phenomenon (*oa-to-ue substitution*) was unique to a single dialect (i.e., Biscayan), and even co-occurrence patterns were limited to a small number of cases, one pattern for Biscayan (i.e., *e-to-i substitution* and *intervocalic t deletion*) and two for Gipuzkoan (*intervocalic g deletion* and *intervocalic r deletion*, and *intervocalic g deletion* and *ea-to-ia substitution*, and *ei-to-i substitution*). From the perspective of ASR, this implies that dialect-aware modeling strategies based solely on dialect labels are unlikely to fully capture the relevant variability. Instead, approaches that explicitly model systematic phonological alternations such as deletion, epenthesis, and substitution patterns may offer a more robust way of improving recognition for non-standard speech. Alternatively, since most ASR systems are trained with data of standard forms of a language, universal speech models such as OpenAI Whisper (Radford et al., 2023) might be more suitable for dialectal speech when dialectal transcriptions are needed, as the transcription of dialectal speech is closer to dialectal transcriptions than standard transcriptions.

The analysis of substitutions further highlights an important methodological issue. Many of the errors contributing to CER arise from vowel and vowel-sequence substitutions (e.g., *ea-to-ia, ua-to-u, oa-to-o*), which are known to be frequent and productive in dialectal Basque (deCastro Arrazola et al., 2015). These alternations often preserve lexical identity for human listeners but create orthographic mismatches for ASR systems evaluated against standardized references. This observation resonates with recent findings in dialectal ASR research showing that the lack of standardized or pronunciation-oriented representations for dialectal speech is a major source of evaluation and modeling difficulties (Vakirtzian et al., 2025). In this respect, our use of pseudo-phonetic transcriptions proves particularly valuable, as it allows us to disentangle errors due to pronunciation variation from those related to lexical errors.

Another relevant aspect of our results concerns the imbalance between dialects. Although Lapurdian, Lower Navarrese and Upper Navarrese are included in the corpus, their limited representation prevents strong statistical conclusions about these varieties, particularly for Lower Navarrese which did not have enough data to be included in our phonological analyses or Souletin that did not have any representation in our dataset. This reflects a broader structural limitation in Basque speech technology resources, where northern dialects remain underrepresented. Consequently, future efforts should prioritize targeted data collection for northern and less represented dialects, as well as for informal and highly spontaneous speech, in order to obtain a more balanced view of dialectal ASR performance.

However, our study has certain limitations that should be taken into account in future research. First, although the standardized transcriptions in our dataset were designed to reflect the orthography of the major dialects included, the considerable degree of subdialectal variation within each Basque dialect, together with the absence of standardized orthography for many varieties, poses challenges for both database annotation and ASR evaluation. Second, the grouping of dialects adopted in this work constitutes another limitation, as there is no broad consensus regarding the number and boundaries of Basque dialects and subdialects (Zuazo, 2008; Aurrekoetxea et al., 2017, 2019). Finally, speech not explicitly identified as dialectal was labeled as Standard Basque, and dialect annotation was performed at the speaker level, even though speakers did not necessarily use dialectal forms consistently. Therefore, we should be parsimonious in the interpretation of our results comparing standard to dialectal Basque.

5. Conclusions

Overall, this study demonstrates that phonological variability in Basque dialects has measurable and systematic effects on ASR performance. The results of our fine-grained phonological analyses provide empirical evidence that supports the need for linguistically informed, dialect-aware modeling strategies in low-resource and typologically distinct languages such as Basque.

6. Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215335 and the project ALIA (Plan Nacional de Tecnologías de Lenguaje-ENIA 2024, SEDIA 19.08.2024), and by a grant from the Department of Culture and Language Policy of the Basque Government (IKER-GAITU project).

7. Bibliographical References

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Gotzon Aurrekoetxea, Ane Arandia, Malalen Camino, and Xantiana Etchebest. 2017. Dialectología perceptual del vasco: aplicación del Little Arrow Method. In Florentino Paredes García, Ana M. Cestero Mancera, and Isabel Molina Martos, editors, *Investigaciones actuales en Lingüística. Vol. V: Sobre variación geolectal y sociolingüística*, pages 51–66. Universidad de Alcalá, Servicio de Publicaciones, Alcalá de Henares, Spain.

Gotzon Aurrekoetxea, Iñaki Gaminde, José Luis Ormaetxea, and Xarles Videgain. 2019. *Euskalkien Sailkapen Berria [New Classification of Basque Dialects]*. UPV/EHU, Bilbao, Spain.

Ekhi Azurmendi, Joseba Fernandez de Landa, Jaione Bengoetxea, Maite Heredia, Julen Etxaniz, Mikel Zubillaga, and Ander Soraluze. 2025. BERnaT: Basque Encoders for Representing Natural Textual Diversity. *arXiv preprint*.

Basque Government. 2016. VI. Inkesta Soziolinguistikoa Euskal Autonomia Erkidegoa [6. The Sociolinguistic Survey of the Basque Autonomous Community]. Technical report, Basque Government, Vitoria-Gasteiz, Spain. Accessed: 2026-02-24.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2025. Lost in variation? evaluating NLI performance in Basque and Spanish geographical variants. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 452–468, Vienna, Austria. Association for Computational Linguistics.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2026. Physical Commonsense Reasoning for Lower-Resourced Languages and Dialects: a Study on Basque. *arXiv preprint*. ArXiv:2602.14812 [cs.CL].

Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R. Mortensen. 2024. Self-supervised Speech Representations Still Struggle with African American Vernacular English. In *Interspeech 2024*, pages 4643–4647.

William Chen, Chutong Meng, Jiatong Shi, Martijn Bartelds, Shih-Heng Wang, Hsiu-Hsuan Wang, Rafael Mosquera, Sara Hincapie, Dan Jurafsky, Antonis Anastasopoulos, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2025. The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties. In *Proceedings of Interspeech 2025*, pages 2093–2097.

Varun deCastro Arrazola, Edoardo Cavirani, Kathrin Linke, and Francesc Torres-Tamarit. 2015. A typological study of vowel interactions in Basque. *Isogloss*, 2015(1/2):147–177.

Ainara Estarrona, Izaskun Etxeberria, Ricardo Etxepare, Manuel Padilla-Moyano, and Ander Soraluze. 2020. Dealing with dialectal variation in the construction of the Basque historical corpus. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 79–89, Barcelona, Spain. International Committee on Computational Linguistics (ICCL).

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECT-BENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. *arXiv preprint*.

Joseba Fernandez de Landa. 2019. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6).

Inma Hernaez, Iker Luengo, Eva Navas, Maren Zubizarreta, Iñaki Gaminde, and Jon Sanchez. 2003. The basque speech_dat (II) database: a description and first test recognition results. In *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 1549–1552.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Ilya Pirkin, and Nishchal Bhandari. 2021. Accented Speech Recognition: A Survey. *arXiv preprint*.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

José Ignacio Hualde. 1991. *Basque Phonology*. Routledge.

José Ignacio Hualde and Iñaki Gaminde. 1998. On the phonology and morphology of Basque dialects. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 32:71–108.

José Ignacio Hualde, Oihana Lujanbio, and Juan Joxe Zubiri. 2010. Goizueta Basque. *Journal of the International Phonetic Association*, 40(1):113–127.

José Ignacio Hualde and Jon Ortiz de Urbina. 2003. *A Grammar of Basque*. Mouton de Gruyter.

José Ignacio Hualde and Koldo Zuazo. 2007. The standardization of the Basque language. *Language Problems and Language Planning*, 31(2):143–168.

Xuedong Huang, Federico Alleva, Hsiao-Wuen Hon, Wei Hwang, and Edward Reichenbach. 2004. Spoken language processing: A guide to theory, algorithm, and system development. *International Journal of Speech Technology*, 7(2):111–126.

JASP Team. 2026. JASP (Version 0.19.1.0) [Computer software]. Accessed: 2026-02-24.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural Language Processing for Dialects of a Language: A survey. *ACM Computing Surveys*, 57(6):149:1–149:37.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):7684–7689.

Saioa Larraza, Arthur G. Samuel, and Miren Lourdes Oñederra. 2017. Where do dialectal effects on speech processing come from? Evidence from a cross-dialect investigation. *Quarterly Journal of Experimental Psychology*, 70(1):92–108.

Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Interspeech 2020*, pages 626–630.

Igor Odriozola, Inma Hernaez, María Inés Torres, Luis Javier Rodriguez-Fuentes, Mikel Peñagarikano, and Eva Navas. 2014. Basque speecon-like and Basque SpeechDat MDB-600: speech databases for the development of ASR technology for Basque. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2658–2665, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mikel Peñagarikano, Amparo Varona, Germán Bordél, and Luis Javier Rodriguez-Fuentes. 2023. Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque–Spanish ASR. *Applied Sciences*, 13(14):848492.

Alec Radford, Jong Wook Wu, Dario Amodei, et al. 2023. Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint. ArXiv:2212.04356 [cs.CL].

Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung yi Lee, and Shinji Watanabe. 2024. ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets. In *Interspeech 2024*, pages 1230–1234.

Rachael Tatman. 2017. "oh, I've Heard That Before": Modelling Own-Dialect Bias After Perceptual Learning by Weighting Training Data. In *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, pages 29–34, Valencia, Spain. Association for Computational Linguistics.

Larraitz Uria and Ricardo Etxepare. 2012. Hizkeren arteko aldakortasun sintaktikoa aztertzeko metodologiaren nondik norakoak: BASYQUE aplikazioa. *Lapurdum. Euskal ikerketen aldizkaria | Revue d'études basques | Revista de estudios vascos | Basque studies review*, (16):117–135.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson. 2018. Joint Modeling of Accents and Acoustics for Multi-Accent Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE Press.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2024. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–614.

Koldo Zuazo. 2008. *Euskalkiak. Euskararen dialektoak*. Elkar.

Systematic Normalization of Spoken Mixed-Language, Mixed-Dialect Data

Margaret Blevins
The University of Texas at Austin
mblevins@utexas.edu

Abstract

Literary transcriptions of spoken language often deviate from standard, written language. These variations can lead to higher than desirable error rates in NLP processing. This is particularly the case for spoken data of low resource varieties, including dialects and contact varieties of higher resource languages. This paper outlines a proposal for the systematic dialect-to-standard normalization of spoken language from language contact and dialect contact situations. This proposed system is then validated and evaluated on the Texas German Sample Corpus (approx. 13 hours), a set of audio and transcripts of Texas German conversations. Texas German is an umbrella term for a set of a heritage varieties of German spoken in Texas, USA that descend from multiple German dialects and that have been in contact with English for 150+ years. The proposed normalization system, along with the accompanying language-tagging system, can act as a starting point for other projects interested in normalizing their mixed variety data.

Keywords: dialect-to-standard normalization, language contact, Texas German

1. Introduction

Transcriptions of spoken language often use modified orthography ("literary transcription", "eye dialect") to more faithfully represent the actual phonetic form of words. In that way, they deviate from standard, written language. At the same time, many NLP tools are trained using standard(-near) texts. When such tools are used on transcriptions of spoken language, it can lead to higher than desirable error rates (see the OOV 'out-of-vocabulary' problem).[1] A similar situation can occur when using NLP tools trained on high resource standard languages on low resource dialects of that language. These issues multiply when working with spoken mixed-language, mixed-dialect data. Creating a normalized annotation layer for transcripts of spoken, mixed-language data can provide automatic NLP tools with more standard-like data, thereby increasing their accuracy.

This paper outlines an approach to normalization for spoken language contact data, mapping non-standard forms to standard language equivalents whenever possible. This system:

- provides a standardized, flexible system to normalize (non-standard, mixed-language, mixed-dialect) spoken data,
- puts normalization decisions on an explicit, linguistically grounded basis, thereby making normalization decisions transparent and reproducible, and
- is publicly accessible.

The importance of this latter point is often overlooked. Although (pre-)processing steps have many implications for searchability and downstream processing and knowing how earlier annotation decisions are made helps researchers use data more accurately and effectively, the decisions made during these steps are often not made public.

As a validation in a real world corpus compilation scenario, the guidelines proposed here are applied to a set of approx. 13 hours of Texas German conversational interview transcripts. Texas German is an umbrella term for a set of heritage contact varieties of German in Texas, USA. As such, it contains a great deal of dialectal variation and contact phenomena, which makes it a useful test for many normalization challenges. The purpose of this study is to develop and test a set of normalization guidelines by manually annotating a set of data and then use those annotations to (a) create a pipeline for automating normalization for this dataset and (b) use the manually annotated data as a gold standard to calculate error rates of the automated process.[2]

[1] E.g., Westpfahl and Schmidt (2013, 140) demonstrate that using an automatic pos-tagger on literary transcriptions leads to a high error rate because the transcriptions of spoken forms do not have lexicon entries in the tagger.

[2] Manually inserting the LANG tags (see section 3) took about 3.3 minutes per audio minute. Manually inserting the NORM tags (see section 5) took about 3.3 minutes per audio minute. Thus, this process took about 7 minutes per audio minute. See conclusion for error rates.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 58–69
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Background

2.1. Texas German

In the mid-1800s, German-speaking immigrants from several dialect regions in central Europe settled in central and south-central Texas (i.e., the "Texas German Belt", see Jordan 1977, 3). Over the last 150+ years, these immigrants and their descendants have come into contact with each other and with other people in the area who speak other languages such as American (Southern) English and (Mexican) Spanish. This resulted in a multitude of (mutually intelligible) contact varieties throughout the Texas German Belt.

Texas German contains a great deal of variation throughout the linguistic spectrum and within its literary transcripts. Section 2.2 below briefly demonstrates several kinds of variation that can appear in transcripts of Texas German audio.[3]

2.2. Variation Within Transcripts

Variation within literary transcripts of spoken language-contact data has three primary sources: the speaker(s), the transcriber(s), and the annotation system(s). Brief examples of each of these within Texas German are provided below.

2.2.1. Inter- and Intra-Speaker Variation

Inter- and intra-speaker[4] variation can occur at all levels of speech and can have dialectical, language contact, or other roots, e.g.,

1. Phonological variation, e.g.,
   [**t**iːɣ̊] vs. [**d**iːɣ̊] 'animal' (SG[5]: [tiːɣ̊])
2. Morphological variation, e.g.,
   er*zählen* vs. ***ver****zählen* 'to tell (e.g., a story)' (SG: *erzählen*)
3. Lexical variation, e.g.,
   Magenschmerzen vs. *Leibweh* vs. *Panzweh* 'stomach ache'
4. Morphosyntactic variation, e.g.,
   (a) case:
   mit ***sie***-NOM/ACC vs. *mit* ***ihr***-DAT 'with her' (SG: *mit ihr*)
   (b) possessive:
   meine Frau ihre Mutter (lit. 'my wife her mother') vs. *die Mutter meiner Frau* (lit. 'the mother my-DAT wife') 'my wife's mother' (SG: *die Mutter meiner Frau*)

Code-mixing and other transfer phenomena can also occur at a variety of levels of speech, e.g.,

5. Foreign language lexeme(s)
   (a) *... wo die* ***fence*** *gebaut ist*
   (TGDP 1-51-1-27-a)[6]
   'where the fence was built / is'
   (SG: *... wo der Zaun gebaut wurde / ist*)
   (b) *sie haben gefragt ob sie — **how do you say 'spies'** — ob sie spies waren*
   'they asked whether they — how do you say 'spies' — whether they were spies'
   (SG: *sie haben gefragt ob sie — wie sagt man 'spies' — ob sie Spione waren*)
6. Mixing of morphemes from different languages
   (a) *ge****jump****t* 'jumped' (SG: *gesprungen*)
   (b) ***fence****n* 'fences' (SG: *Zäune*)
7. Phrasal calques
   ich kann nicht für sicher sagen
   I can not for sure say

   'I can't say for sure.'

   (SG: *ich kann nicht mit Sicherheit sagen*)

2.2.2. Inter- and Intra-Transcriber Variation

Once speakers produce spoken language, how a transcriber depicts that language can add another layer of variation within transcripts. A single transcriber may transcribe something inconsistently, or there may be variation between different transcribers, e.g.,

8. Use of special characters:
   *gro****ß*** vs. *gro****ss*** 'big' (SG: *groß*)
9. Use of abbreviations and/or punctuation:
   St. *John****'s*** vs. ***Saint*** *John****s***
10. Location and inclusion of word boundaries:
    nennses *"links"* vs. ***nen se s*** *"links"*
    (*nennen sie es "links"*, 'they call it "links"',

[3] For the purposes of this study, I use literary transcripts from the Texas German Dialect Project Corpus (TGDP Corpus). This corpus is part of the larger Texas German Dialect Project (TGDP, tgdp.org).

[4] Intra-speaker variation refers to variation within a single speaker's speech—a single person may vary, e.g., how they say a particular word throughout a conversation. Inter-speaker variation refers to variation between different people, e.g., person A and person B may use different lexemes for a particular object.

[5] SG = Standard German

[6] This is an interview segment ID number from the TGDP Corpus. It can be read as follows: the first number is the interviewer ID, the second number is the speaker ID, the third number is the interview number, and the fourth number is the interview segment number. Thus, this is the twenty-seventh segment of interview number 1 that occurred between interviewer 1 and speaker 51.

lit. 'call they it "links"')
(SG: *nennt man sie Verbindungen*)

11. Difference in identification/interpretation:
[foːɐ̯] being transcribed as *for* (English) vs.
vor (German)

12. Transcriber misunderstanding:
poodles (English) vs. *Puter* 'turkeys'

Inter-speaker and inter-transcriber variation are present in any transcription of spoken language. Transcriber variation can be reduced by having detailed transcription guidelines and transcription quality checks.[7]

2.2.3. Inter-Corpus and Transcription System Variation

Additionally, there is inter-corpus variation with respect to transcription systems. There are many systems for transcribing spoken language, e.g., cGAT (Schmidt et al., 2023), HIAT (Rehbein et al., 2004), and Jefferson (Jefferson, 2004). Example (13) shows how a reference to the Texas-based grocery store H-E-B could be transcribed according to two transcription systems:

13. Transcriptions referring to H-E-B

    (a) cGAT (Schmidt et al., 2023, 33)
    ‹aitch› ‹ey› ‹bee›

    (b) TGDP project-internal guidelines
    ‹H-E-B› (or) ‹HEB› (or) ‹H.E.B.›

The two transcriptions above differ both in terms of orthography (i.e., which characters are used to represent each letter of the alphabet) and in terms of how many tokens they contain (three tokens vs. one token, respectively).

While researchers may want to minimize some variation (e.g., transcriber variation), they may want to embrace other kinds of variation to achieve a project's goals (e.g., dialectal variation). Deciding what to (not) include within transcriptions and annotations is an integral step of designing and building a corpus.

[7]The TGDP Corpus transcriptions are particularly variable due to (a) the amount of speaker variation, (b) the project's relatively limited transcription guidelines, and (c) the fact that L1, L2, and L3 German speakers transcribe the data, none of whom are native speakers of Texas German and all of whom have varying degrees of familiarity with German dialects. For example, the German word *wahrscheinlich* 'probably' is spelled in at least 10 different ways: *wahrscheinlich, verscheinlich, washeinlich, wahrscheinlih, wahrscheinlich, vescheinlich, vescheilich, warscheinlich, warscheinlih*, and *werscheinlich*. See also Bailey et al. (2005) for a discussion of transcriber effects on phonetic transcription.

2.3. Definitions of Normalization

As mentioned in the introduction, this paper proposes a systematic normalization system. Normalization is a common pre-processing step in NLP which systematically decreases variation within data, thereby making downstream processing more accurate.

CLARIN describes normalization as "the process of transforming parts of a text into a single canonical form" (CLARIN). However, which canonical form to choose (and why) differs between researchers and research communities.[8]

Within the NLP community, normalizing often means something like reducing all characters to their lower-case equivalents, **stemming**, and/or **lemmatization** (Bird et al., 2009, 107–108).[9] Another normalization task can be "identifying non-standard words, including numbers, abbreviations, and dates, and mapping any such tokens to a special vocabulary" (Bird et al., 2009, 108).[10] For example, ‹July 4, 2020›, ‹4th of July, 2020›, and ‹Fourth of July, 2020› could all be normalized to ‹2020-07-04›. Text normalization can also refer to "the task of transforming lexically variant words to their canonical forms" (Mehmood et al., 2020).[11]

Similarly, in the computer-mediated communication (CMC) community, the term text normalization has been used to mean "the challenge of discovering the English [i.e., the relevant standard language] words corresponding to the unusually-spelled words used in social-media messages and posts" (Jahjah et al., 2016, 180).[12]

Within the Text-to-Speech research community (TTS),[13] text normalization means "convert[ing] a written representation of a text into a representation of how that text is to be read aloud" (Yolchuyeva et al., 2018, 589), e.g., changing ‹$5› to ‹five dollars› (Baumann et al., 2019).

The term normalization has also been used within the historical corpus community, albeit with differing meanings. Bollmann et al.'s (2012) corpus of Early New High German texts contains both

[8]Bigi (2014) provides a brief overview of previous approaches to text normalization (a task that is generally thought to be language-dependent and task-dependent) and attempts to make a text normalization approach that is as little language- and task-dependent as possible.

[9]Also sometimes referred to as **text normalization**. Both stemming and lemmatization have the goal "to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form" (Manning et al., 2009, 32). For information, see Manning et al. (2009, 32–34).

[10]See also Sproat et al. (2001, 289), Bigi (2014, 515).

[11]See also **lexical normalization** (Toska, 2020).

[12]See also Kuparinen (2023).

[13]Often linked to the Automatic Speech Recognition (ASR) community.

a normalization and a **modernization** annotation layer. The normalization layer "stay[s] close to the original and, e.g., keep[s] historical inflection even if it violates modern morpho-syntactic constraints" while the modernization layer "adjust[s] inflection to [fit] modern constraints" (Bollmann et al., 2012, 342).[14] There are several corpora that refer to these kinds of annotation processes as **regularization** or regularized spelling, e.g., the EarlyPrint project and the British National Corpus (BNC).[15] Alternatively, Jurish et al. (2013) describe **canonicalization** as "assigning an extant equivalent to each word of the input text and deferring application analysis to these canonical cognates" (see also Jurish (2010, 2011); Gotscharek et al. (2009); Reffle et al. (2008); Makarov and Clematide (2020)).

A similar approach, **dialect-to-standard normalization**, i.e., "transforming dialectal text into a standard variety while maintaining as much of the original meaning as possible," has been used to help lower-resource varieties benefit from the tools built for higher-resource standardized varieties (Dimakis et al. (2025, 1); see also Joshi et al. (2025)).[16]

It suffices to say that the term normalization does not have a single, consistent definition across corpora and different sub-fields of linguistics. For the purposes of this paper, I use the term normalization to refer to mapping literary transcription forms to their standard-orthographic equivalent (e.g., *zwohundert* → *zweihundert* 'two hundred').[17] This definition does not state, however, that the single representation needs to be a standard, *grammatical* language form, see section 5 for more information.

[14]The RIDGES (Register in Diachronic German Science) corpus (Lüdeling et al., 2020) has similar annotation layers but calls them by different names: RIDGE's 'clean' annotation layer appears to be comparable to Bollman et al.'s (2012) 'normalization' layer, and RIDGE's 'normalization' layer appears to be similar to Bollman et al.'s (2012) 'modernization' layer.

[15]See Burnard (2007, 324) for information about regularized spelling in the BNC.

[16]For more information about dialect-to-standard normalization, see e.g., Kuparinen et al. (2023), Partanen et al. (2019).

[17]The term **normorthographic** is also used in similar corpora, such as the SiNProjekt (Sprachvariation in Norddeutschland, 'Language Variation in Northern Germany') and Regionalsprache.de (REDE) (Kehrein and Vorberger, 2018, 141–142). As Rehbein and Schalowski (2013, 208) note, "[a]dding an orthographic normalisation to the transcription might be seen as a 'poor man's target hypothesis.'"

2.4. Previous Corpora

The guidelines outlined in sections 3-5 below are based on the annotation guidelines for 10 non-standard varieties of spoken German, five of which are extraterritorial contact varieties: the Texas German Dialect Project (TGDP) (Boas et al., 2010); the *Namdeutsch* 'Namibia German' corpus (DNam) (Zimmer et al., 2020); the *Unserdeutsch* (Rabaul Creole German) corpus (Maitz et al., 2016; Maitz and Volker, 2017; Götze et al., 2017); the *Russlanddeutsche Dialektdatenbank* 'Russian German Dialect Database' (RuDiDat) (Berend and Frick, 2021); and the *Digitales Portal: Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus Projekt* (UZSK) 'Digital portal: Bilingual and language contact corpus of German as a minority language in Hungary' (Földes, 2016a,b,c).

The other five corpora are spoken varieties of (non-standard) German within DACH: the *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FOLK), a general corpus of spoken German (Kaiser, 2018; Schmidt, 2016, 2018; Deppermann and Hartung, 2011); the *KiezDeutsch* corpus (KiDKo), a corpus of speech from relatively mono- and multi-ethnic communities in Berlin (Wiese et al., 2012; Rehbein et al., 2004); the Berlin Map Task corpus (BeMaTaC), a learner corpus (Sauer and Lüdeling, 2016); ArchiMob, a dialectal, Swiss German corpus (Scherrer et al., 2019a,b); and a code-switching corpus of Turkish-German (SAGT) (Çetinoglu, 2017).[18]

3. Language Tagging

The first step of normalizing language contact data is to language tag the data. This is because the language assigned to a token has implications for how it should be normalized. For example, if the token <ham> (i.e., a common spoken form of German *haben* 'to have') were tagged as German, then it should normalized to <haben> (i.e., the standard German form) but if the same token were tagged as English, it should be normalized as <ham> (meaning 'pork meat product'). This language interpretation also has implications for other annotations such as part-of-speech, with <haben> being tagged as a verb and <ham> being tagged as a noun.

The LANG annotations in the system proposed in this paper are token annotations in which the annotator records which language(s) they interpret a particular token to be.[19] The purpose of the language layer is to make annotators' interpretations and assumptions transparent. For example, this

[18]See language resources section for corpora URLs.

[19]The language tags are based on the ISO 639-2 guidelines (ISO).

fictional Texas German sentence would be tagged as follows:

14. **tok** mir ham die die gerollt
    lang deu deu deu eng mix.deu+eng

Example (14) means *Wir haben den Würfel geworfen* 'We rolled the die.' These language tags are based on each token's phonological form, morphological form, part-of-speech, semantics, and (sometimes) surrounding tokens.

As a first step, if a token's (phonological[20] and morphological) form, meaning, and part-of-speech all correlate to a lexical entry in a standard reference work, then it should be tagged with a standard language tag, such as `deu` (German) or `eng` (English), see Figure 1.

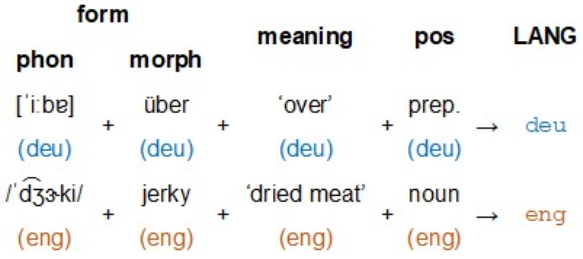


Figure 1: Depiction of scenarios leading to standard language tags.

When a token's form, meaning, and/or part-of-speech do *not* correlate to a lexical entry in a standard reference work, the token should be language-tagged with something other than a standard language tag. This can occur for several reasons, briefly illustrated below.

If a token has a form that is made up of German morphemes but there is no lexical entry for that token's intended meaning and/or part-of-speech, then it should be tagged as `deu.txg`. See e.g., Figure 2.[21]

If there is morphological/lexical material from multiple languages present in a single token, that token should be marked as `mix.LANG+LANG`. For example, the token *gemoved* 'moved' would be tagged as `mix.deu+eng` because it is comprised of a mixture of German *ge-* and English *moved*. The language tags are written in alphabetical order

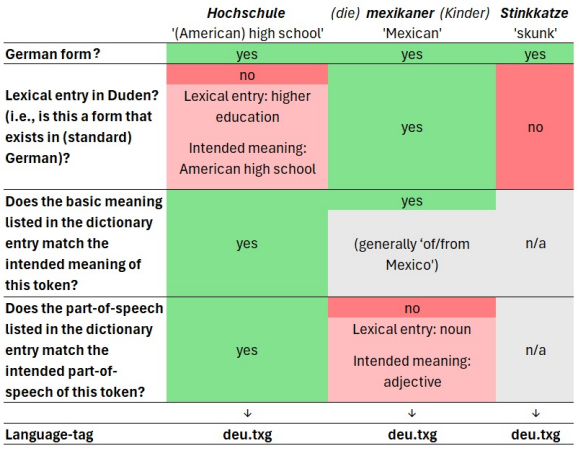


Figure 2: Different `deu.txg` scenarios.

to keep them consistent throughout the corpus.[22]

The language-tag `xxx` is used in situations in which it is unclear which language a particular token is. For example, all material marked as unintelligible (indicated by a (???) in the TOK layer) would be language-tagged as `xxx`.[23]

If a token was incomplete (e.g., the speaker cuts themselves off mid-word), but the presumed language of the intended token was relatively unambiguous, then the token should be language-tagged with the appropriate language, albeit with a hash tag (#) immediately preceded the language token to mark it as an interpretation of an incomplete utterance.

An asterisk is used to indicate that there are multiple possible interpretations for a given token. Each interpretation is marked with an asterisk. For example, the token *mädchens* could have been interpreted as an entirely German token, albeit with a nonstandard German plural marker (the standard German plural would be the null-plural *die Mädchen*), or it could have been interpreted as a German noun (*Mädchen*) + the English plural marker *-s*. The former interpretation would lead to a `deu` language-tag, while the later interpretation would lead to a `mix.deu+eng` language-tag. Both of these interpretations are included in the language-

[20] The phonology does not need to be a perfect match, but it should be similar.

[21] If one wanted to annotate for several dialects under the umbrella of a standard language, one could use different subtags, e.g., `deu.bai` for Bavarian or `deu.pla` for Low German. The main thing is to systematically and consistently use a set of language tags throughout an entire dataset. Distinguishing between related dialects is its own challenge, see e.g., Heeringa (2004).

[22] Pronouncing English words with a German "accent" does not trigger the use of the `mix` tag.

[23] In the version of these guidelines described in Blevins (2022), the `xxx` tag includes three subtags for material that is consistently language-ambiguous: `xxx.hes` (filled pauses / vocalized hesitations), `xxx.exc` (the exclamation *oh*), and `xxx.bcl` (backchanneling, such as *uh-huh* to indicate agreement). This was done to expedite the manual annotation process. Including labels to common non-lexical utterances such as `xxx.hes` helped improve subsequent pos-tagging.

tag layer, e.g.,[24]

15. **tok** mädchens
    lang *deu *mix.deu+eng

For a summary of the language tags and symbols within the LANG layer, see Table 1.[25]

4. Ambiguous Situations withing Language Tagging

I approach potentially ambiguous tokens in one of three ways: create a system to disambiguate them, always mark them as ambiguous, or always mark them in the same way, although they are ambiguous. These approaches are discussed in more detail below.

4.1. Disambiguating Cognates and Homographs Using Phonology

One way of disambiguating potentially language-ambiguous tokens is to rely on phonology. This is particularly useful if the pronunciation of the tokens is question is quite distinct. For example, German *Musik* [muˈziːk] has second syllable stress while English *music* /ˈmjuzɪk/ has first syllable stress, making them usually easy to differentiate.

4.2. 4.2 Cognates and Homographs with Similar Phonology

For cognates that are more similar than those described in 4.1, there are three options: (1) always mark such tokens as ambiguous, (2) always choose the same language for a particular token, or (3) have a system for deciding whether to mark something as a single language or ambiguous between multiple languages. All three of these approaches are taken in different situations within this proposed annotation system.

[24]The asterisk was used in a similar way in the normalization layer to indicate ambiguity.

[25]The standard language reference works used are as follows: German = *Duden* (online, www.duden.de); English = Merriam-Webster online (www.merriam-webser.com); Spanish = *Diccionario del español de México* 'Dictionary of Mexican Spanish' (DEM) (https://dem.colmex.mx//); Czech = (1) *Slovník spisovného jazyka českého* 'The Dictionary of the Written Czech Language' (2011), https://lexiko.ujc.cas.cz/texts/ssjc.html, (2) ORAL2013 portion of the Czech National Corpus (www.korpus.cz) (Křen, 2015; Válková et al., 2012), and (3) *Deutsch-böhmisches Wörterbuch* (Dobrovský et al., 1802-1821); Wendish = (1) *Prawopisny słownik hornjoserbskeje rěče: Hornjoserbsko-němski słownik* 'German-Upper Sorbian Dictionary' (Völkel, 2005), (2) *Niedersorbisch-deutsches Wörterbuch* 'Lower Sorbian-German Dictionary' (Starosta, 1999) (niedersorbisch.de).

4.2.1. Tokens That are Always Marked as Ambiguous

For this particular language pair (German-English), the following situations were always marked as language ambiguous.

Situation 1: A speaker pluralizes a German noun with an *-s* plural marker (for German nouns that use plural markers other than *-s* in standard German), e.g., *Tante**s*** 'aunts' (SG: *Tante**n***), *Baum**s*** 'trees' (SG: *B**ä**um**e***). This is because the *-s* plural marker is present in both German and English. Therefore, if it occurs in an "unexpected" place, it is unclear if it is a non-standard German plural ending or a borrowing of an English plural ending.

Situation 2: Tokens for which it is unclear whether a speaker is using an English loan that has been phonologically adapted to reflect German-like pronunciation, or whether they are using a term that was in the original Texas German donor dialects, and that that term has gone through semantic shift (e.g., *Grad* '(school) grade' (SG: *Klasse*), *Trubel* 'trouble' (SG: *Ärger*).

Situation 3: When it is ambiguous whether a speaker omitted a final schwa or whether they pronounced an English word according to German phonology, e.g., *Seit* 'side' (SG: *Seite*).

4.2.2. Ambiguous Tokens That are Always Tagged as a Single Language

There are two particularly common tokens that are always language-tagged as `deu` although they could technically be considered ambiguous.

First, Texas German speakers commonly say *denn* meaning 'then' (standard German *dann*). It is ambiguous whether *denn* should have the language tag `deu` (implying that it is a dialectal, phonological variant of *dann*, leading to the normalization form *dann*), `deu.txg` (implying a that it is German form with a non-standard meaning, which would lead to the normalization form *denn*), or `eng` (implying that it is a dialectal, phonological form of English *then*, leading to the normalization form *then*). Thus, if one wanted to capture all three of these potential interpretations, one would language-tag (and normalize) *denn* 'then' as `*deu *deu.txg *eng` (leading to the normalization forms ***dann* *denn* and ***then*, respectively). However, in the proposed system, *denn* 'then' is language-tagged as `deu` and normalized as <dann>. This is done because *denn* is an attested form of standard German *dann* 'then' in continental German dialects (e.g., in Cologne[26]).

[26]https://dat-portal.lvr.de/woerterbuch/D – "denn": "Häufig für 'dann': "Gestern wa ich ers bein Aazt, un denn wa ich noch auf en Maakt. Wenn schon, denn schon! (nicht kleckern, sondern klotzen!) Ich kauf mir

Tag	Explanation	Example
`deu`	German	*Kuh, gespielt*
`deu.txg`	German material ... (a) with non-standard semantics, or (b) that does not appear in *Duden*	(a) *Luftschiff* 'airplane', *Hochschule* '(American) high school' (b)*Stinkkatze* 'skunk', *mitaus* 'without'
`eng`	English	*well, baseball*
`spa`	Spanish	*leche* 'milk'
`ces`	Czech	*pivo* 'bier'
`wen`	Wendish	*wótnic* 'grandmother'
`mix.lang + lang` (e.g., `mix.deu+eng`)	Mixture of morphological / lexical material from multiple languages	*Hickorybaum* 'hickory tree', *gemovt* 'moved', *schlipperich* 'slippery'
`* + lang` (e.g., `*deu`)	Ambiguous language	*in* 'in', *Mädchens* 'girls'
`# + lang` (e.g., `#deu`)	Unknown but (partially) reconstructable material	*ausgezeich*
`xxx`	Unintelligible material, material for which the intended language is indeterminate	(???), *g*

Table 1: List of available language tags

A similar approach is taken with the token *is* 'is'. If a speaker pronounces *is* as [ɪs] (not US-American English /ɪz/) and if the token is otherwise surrounded by German tokens, then it is marked as `deu` and not as ambiguous.

16. **tok** sie is schnell
lang deu deu deu
norm sie ist schnell
'she is fast'

Similar to *denn* 'then,' this is done because *is* [ɪs] is an attested colloquial pronunciation of standard German *ist* [ɪst] and practically speaking, this saves time for the annotator, as this is a common occurrence in the corpus. Otherwise, it is treated like *in*, see section 4.2.3.

4.2.3. Tokens for Which There is a Disambiguation System

Certain tokens have a great deal of overlap with regards to orthographical form, meaning, and pronunciation in English and German. One common example of that is the preposition *in*.

Instead of always tagging *in* as `deu`, `eng`, or `*deu *eng`, the language tag is reliant on the language of the tokens to the immediate left and right of the in token. If *in* is surrounded by German tokens, it is marked as `deu` (e.g., (17)). If it is surrounded by English tokens, it is tagged as `eng` (e.g., (18)). If a German token is on one side, and an English token on the other, it is tagged as ambiguous (e.g., (19)). If it is surrounded by German tokens but used in a way that is not available in standard German, it is tagged as `deu.txg` (e.g., (20)).

17. **tok** irgendwo in die gegend
lang deu deu deu deu
'somewhere in the area'
(TGDP 1-51-1-9-a)

18. **tok** fighting in the civil war
lang eng eng eng eng eng
(TGDP 1-51-1-23-a)

19. **tok** nicht mehr in saloon
lang deu deu *deu *eng eng
'[she can] no longer [go] into the saloon'
(TGDP 10-171-3-24-a)

20. **tok** was das in deutsch ist
lang deu deu deu.txg deu deu
'... what that is in German'
(TGDP 9-122-5-1-13-a)

5. Normalization

5.1. Definition

The NORM ('normalization') layer is a token annotation that exists in addition to the above-mentioned layers. I use the term *normalization* to mean associating transcribed utterances to a set of consistent, systematic representations, the majority of which follow standard language orthography. This can be seen as a kind of dialect-to-standard normalization. I am primarily guided by Scherrer and Samardžić (2016, 1) idea of "mapping the variants of what can be identified as the

jetzt nen Daimler, wenn schon, denn schon!"

same word to a single representation," Bird et al.'s (2009, 108) notion of "identifying non-standard words […] and mapping any such tokens to a special vocabulary," and the desire to map each token to the closest standard language equivalent while preserving the intended lemma and part-of-speech.

The normalizations in this proposed system are token-based reductions of phonological and/or orthographic variation. No changes are made to grammar (e.g., case, gender), (lexical / morphological) word choice, word order, part-of-speech, or lemma. No tokens are inserted or removed, but contractions can be split into their constituent parts.

For example, normalization can disambiguate homographs that occur due to Texas German pronunciation patterns, e.g., both *Tür* 'door' and *Tier* 'animal' being pronounced (and transcribed) as ‹tier› due to the presence of an unrounded vowel where standard German has a rounded vowel. In the process of normalization, ‹tier› 'animal' would be normalized to ‹Tier› and ‹tier› 'door' should be normalized to ‹Tür›, thereby disambiguating these instances.

These normalizations are not a translation into (more) standard German, nor are they a "correction" of non-standard grammatical variation. For example, *Schrippe* 'roll/bun' (Berlin dialect) would not be translated to the more ubiquitous, standard term *Brötchen*, as these are two entirely different lexemes. Also, if a speaker says *mit er*-NOM or *mit ihn*-ACC 'with him' although standard German would require the dative *ihm* for 'him' following the preposition *mit* 'with', *mit er* and *mit ihn* would be normalized to *mit er* and *mit ihn* (i.e., no changes would be made to their forms) because it would be inaccurate to claim that *er* and *ihn* are phonological or orthographic forms of *ihm*.[27]

There are several main principles that these normalization guidelines follow:

1. Normalization forms are based on the presumptions implied by the language tag each token is given. (e.g., see ‹ham› discussion in section 3).

2. If a token is marked with a standard language language-tag (e.g., `eng` or `deu`), then the orthographic form in the normalization layer should be an existing standard language reference work.[28]

3. The presumed part-of-speech and lemma of the token in the TOK layer should also be the same as the part-of-speech and lemma for the token in the NORM layer

4. In order to achieve this form (as described in points 2 and 3), phonological and orthographic changes are permissible, while lexical and grammatical changes are not.

5. Tokens language-tagged as something other than a standard language have different rules, see section 5.2.

5.2. Non-Standard Situations

Tokens tagged as `mix` or `deu.txg` were morphologically normalized, i.e., each morphological segment of a token was normalized in its own right, based on each morpheme's presumed source language if there was one, see Table 2.

LANG	transcribed form	NORM
`deu.txg`	‹Stinkkatz› *stink* 'stink' (deu) + *Katze* 'cat' (deu)	‹Stinkktze› 'skunk'
`mix.deu+eng`	‹Futtershtore› (*Futter* 'feed' (deu) + *shtore* (eng) 'store'	‹Futterstore› 'feed store'
	‹fencen› *fence* 'fence' (eng) + *n* (plural) (deu)	‹fencen› 'fences'
	‹schlipperich› *slipper-* (eng) + *-ig* 'y' (deu)	‹slipperig› 'slippery'

Table 2: Normalization of tokens with non-standard tags

If a token was incomplete (e.g., the speaker cuts themselves off mid-word), but the presumed intended word was relatively unambiguous (e.g., due to a later mention of the same word) then the normalized token was preceded by a pound sign (`#`) to mark it as a reconstruction of an incomplete utterance. Portions (syllables) of the reconstructed utterance that were not produced by the speaker

[27] This differs from, say, an L2 corpus of German in which annotators may want to annotate speaker "errors" / target hypotheses. Although Texas German shares many similarities with L2 German, it is not the goal of this annotation to make their utterances conform to standard German grammar.

[28] For each standard language tag, a standard reference work was chosen to guide the spelling of each normalized form. The reference work(s) a given project uses should be explicitly stated. See footnote 25 for a list of the standard reference works used in this study. While this is possible for (varieties of) standardized languages, a slightly different approach may be necessary when normalizing a (variety of a) language that does not have a standardized orthography. For example, if one wanted to normalize a corpus of Pennsylvania German, it would be helpful to choose a particular dictionary / spelling convention to guide the normalization process and make those decisions public.

originally were included in square brackets, e.g., <#ausgezeich[net] > 'excel[lent]'.

Backchanneling utterances contained a great deal of orthographic variation. In order to systematize the forms, each meaning of a backchanneling signal was given a single form. For example, all affirmative signals such as <hmhm>, <uh-huh>, and <mhmm> were normalized to the form <hmhm> (see Blevins, 2022, 627).

6. Texas German Sample Corpus

In order to develop and test the effectiveness of the language-tagging and normalization guidelines as described above, I manually annotated a set of Texas German conversational audio (the Texas German Sample Corpus, TGSC). The audio and initial transcripts of the TGSC come from the Texas German Dialect Project Corpus, a collection of Texas German interviews from 2001-today (see Boas et al. (in press)).

The TGSC is collection of randomly selected audio from 150 randomly sampled speakers that is representative of the TGDP's first 600 Texas German speakers who:

- were recorded by the TGDP itself (i.e., not recordings that were subsequently donated to the project),
- do *not* speak a particular unique sub-dialect of Texas German (e.g., Texas Alsatian),
- were born in Texas and still live in Texas,
- whose ancestors came to Texas before 1917, and
- grew up speaking Texas German and learned Texas German at home through relatives.

The TGSC is proportionally representative based on the extra-linguistic variables sex and birthplace (county) and is available for download at https://doi.org/10.18738/T8/IOX9ZA.

7. Conclusion

The guidelines outlined here are presented in order to highlight some of the challenges that occur when normalizing spoken language-contact data and proposes possible solutions to those challenges. Admittedly, teasing apart two related languages (here: German and English) may have different challenges than other contact situations (e.g., a tonal language in contact with an atonal language, or a language in contact with another language that uses a different writing system, etc.). These guidelines are therefore not intended as a one-size-fits all solution; they provide a resource for other projects to use as a starting point, and/or to react against rather than starting from scratch. It is a step towards transparency of pre-processing decisions, thereby facilitating downstream processing and the accuracy of future research using these materials.

After the TGSC was constructed, it was used to train an automatic tagger to add LANG and NORM annotations to the larger TGDP Corpus. Following the process described in Boas et al. (in press), this led to a language tagging error rate of 6% and a normalization error rate of 7%. One of the additional benefits of the TGSC itself is that it is a step towards developing a comprehensive dictionary of (dialectal) terms that do not exist in any of the related standard language dictionaries (cf. Dimakis et al., 2025).

8. Acknowledgments

The author would like to thank Thomas Schmidt for his valuable insights and support; Hans C. Boas, the founder and director of the Texas German Dialect Project, for his support and for providing access to the data; and the hundreds of Texas German speakers who have generously agreed to be interviewed by the TGDP.

9. Availability

The Texas German Sample Corpus is available for download at https://doi.org/10.18738/T8/IOX9ZA and may be used for non-commercial and non-AI/LLM purposes. See https://tgdp.org/dialect-archive/ for more information regarding user rights and responsibilities.

10. Ethics Statement

All participants were informed about the project and the intended use of the recordings. Written consent was obtained for recording, transcription, and inclusion of their data in the corpus. Directly personally identifying information was removed during transcription. When making the data available, I am complying with FAIR principles, observing the restrictions resulting from the informed consent.

11. Bibliographical References

G. Bailey, J. Tillery, and G. Andreas. 2005. Some effects of transcribers on data in dialectology. *American Speech*, 80(1):3–21.

T. Baumann, A. Köhn, and F. Hennig. 2019. The Spoken Wikipedia Corpus: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation*, 53(2):303–329.

N. Berend and E. Frick. 2021. Russlanddeutsche Dialekte online: Dokumentation, Präsentation und Recherche deutscher Auslandsvarietäten im Internet am Beispiel des Russlanddeutschen. In C. Földes, editor, *Kontaktvarietäten des Deutschen im Ausland*. Narr Francke Attempto.

B. Bigi. 2014. A multilingual text normalization approach. In *Human language technology challenges for computer science and linguistics, 8387*, pages 515–526.

S. Bird, E. Klein, and E. Loper. 2009. *Natural language processing with Python*, 1st edition. O'Reilly.

M. Blevins. 2022. *The Language-Tagging and Orthographic Normalization of Spoken Mixed-Language Data, with a Focus on Texas German*. Ph.D. thesis, The University of Texas at Austin.

H.C. Boas, M. Pierce, K.A. Roesch, G. Halder, and H. Weilbacher. 2010. The Texas German Dialect Archive: A multimedia resource for research, teaching, and outreach. *Journal of Germanic Linguistics*, 22(3):277–296.

H.C. Boas, T. Schmidt, and M. Blevins. in press. A new corpus platform for the Texas German Dialect Project. *Language Resources and Evaluation*. To appear.

M. Bollmann, S. Dipper, J. Krasselt, and F. Petran. 2012. Manual and semi-automatic normalization of historical spelling: Case studies from Early New High German. In *Proceedings of KONVENS 2012, The 11th Conference on Natural Language Processing (LThist 2012 workshop)*, Vienna, Austria.

L. Burnard, editor. 2007. *Reference Guide for the British National Corpus (XML Edition)*.

CLARIN. *Tools for normalization*.

Ö. Çetinoglu. 2017. A code-switching corpus of Turkish-German conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*, Valencia, Spain.

A. Deppermann and M. Hartung. 2011. Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des 'Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) am Institut für Deutsche Sprache (Mannheim). In M. Müller E. Felder and F.Vogel, editors, *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen*, pages 414–450. De Gruyter.

A. Dimakis, J. Pavlopoulos, and A. Anastasopoulos. 2025. Dialect normalization using large language models and morphological rules. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 23696–23714.

J. Dobrovský, S. Leška, A. J. Puchmajer, and V. Hanka. 1802-1821. *Deutsch-böhmisches Wörterbuch*. In der Herrlischen Buchandlung, Prag.

C. Földes. 2016a. Diskurse im Wirkungsraum von Zweisprachigkeit: Werkstattbericht aus einem Forschungs- und Dokumentationsprojekt. In M. Dus, R. Kolodziej, and T. Rojek, editors, *Wort–Text–Diskurs*, page 321–336. Peter Lang GmbH Internationaler Verlag der Wissenschaften.

C. Földes. 2016b. Ungarndeutsche Sprachvariation und Mehrsprachigkeit: Ein Korpusprojekt auf der Basis von empirischer Feldforschung und Online-Sprachdokumentation. *Sprachtheorie und germanistische Linguistik*, 26(2):167–190.

C. Földes. 2016c. Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus: Konzept, Design und Inhalte. *Zeitschrift für interkulturelle Germanistik*, 7(1):167–181.

A. Gotscharek, U. Reffle, C. Ringlstetter, and K. U. Schulz. 2009. On lexical resources for digitization of historical documents. In *Proceedings of the 9th ACM Symposium on Document Engineering*, pages 193–200, New York. Association for Computing Machinery.

A. Götze, S. Lindenfelser, S. Lipfert, K. Neumeier, W. König, and P. Maitz. 2017. Documenting Unserdeutsch (Rabaul Creole German): A workshop report. *Journal of the Linguistic Society of Papua New Guinea*, pages 65–90.

W. J. Heeringa. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University of Groningen.

ISO. 2016. *ISO 24624:2016 Language resource management — Transcription of spoken language*.

V. Jahjah, R. Khoury, and L. Lamontagne. 2016. Word normalization using phonetic signatures. In *Proceedings of the 29th Canadian Conference on Artificial Intelligence*, pages 180–185.

G. Jefferson. 2004. Glossary of transcript symbols with an introduction. In G. H. Lerner, editor, *Conversation Analysis: Studies from the First Generation*, pages 13–31. John Benjamins, Amsterdam and Philadelphia.

T. G. Jordan. 1977. The German element in Texas: An overview. *Rice University Studies*, 63(3):1–11.

A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, and D. Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

B. Jurish. 2010. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, page 72–77, Uppsala, Sweden. Association for Computational Linguistics.

B. Jurish. 2011. *Finite-state canonicalization techniques for historical German*. Ph.D. thesis, Universität Potsdam.

B. Jurish, M. Drotschmann, and H. Ast. 2013. Constructing a canonicalized corpus of historical german by text alignment. In P. Bennett, M. Durrell, S. Scheible, and R. J. Whitt, editors, *New Methods in Historical Corpora*, pages 221–234. Narr.

J. Kaiser. 2018. Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 19:515–552.

R. Kehrein and L. Vorberger. 2018. Dialekt- und Variationskorpora. In M. Kupietz and T. Schmidt, editors, *Korpuslinguistik*, pages 125–150. De Gruyter.

M. Křen. 2015. Recent developments in the Czech National Corpus. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 1–4.

O. Kuparinen. 2023. Murrevikko – a dialectally annotated and normalized dataset of Finnish tweets. In *Proceedings of the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 31–39, Dubrovnik. Association for Computational Linguistics.

O. Kuparinen, A. Miletić, and Y. Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

A. Lüdeling, C. Odebrecht, Schnelle G. Krause, T., and C. Fischer. 2020. Ridges herbology.

P. Maitz, W. König, and C. A. Volker. 2016. Unserdeutsch (Rabaul Creole German): Dokumentation einer stark gefährdeten Kreolsprache in Papua-Neuguinea. *Zeitschrift für germanistische Linguistik*, 44(1):93–96.

P. Maitz and C. A. Volker. 2017. Documenting Unserdeutsch: Reversing colonial amnesia. *Journal of Pidgin and Creole Languages*, 32(2):365–397.

P. Makarov and S. Clematide. 2020. Semi-supervised contextual historical text normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7284–7295.

C. D. Manning, P. Raghavan, and H. Schütze. 2009. *An introduction to information retrieval*. Cambridge UP.

K. Mehmood, D. Essam, K. Shafi, and M. K. Malik. 2020. An unsupervised lexical normalization for Roam Hindi and Urdu sentiment analysis. *Information Processing & Management*, 57(6).

N. Partanen, M. Hämäläinen, and K. Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT)*, pages 141–146, Dubrovnik, Croatia.

U. Reffle, C. Ringlstetter, A. Gotscharek, and K. U. Schulz. 2008. Successfully detecting and correcting false friends using channel profiles. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):17–22.

I. Rehbein and S. Schalowski. 2013. STTS goes Kiez: Experiments on annotating and tagging urban youth language. *Journal for Language Technology and Computational Linguistics*, 28(1).

J. Rehbein, T. Schmidt, B. Meyer, F. Watzke, and A. Herkenrath. 2004. *Handbuch für das computergestützte Transkribieren nach HIAT*, volume 56 of *Arbeiten zur Mehrsprachigkeit*.

S. Sauer and A. Lüdeling. 2016. Flexible multilayer spoken dialogue corpora. *International Journal of Corpus Linguistics, Special issue: Complation, transcription, markup and annotation of spoken corpora*, 21(3):419–438.

Y. Scherrer and N. Samardžić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.

Y. Scherrer, T. Samardžić, and E. Glaser. 2019a. ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454.

Y. Scherrer, T. Samardžić, and E. Glaser. 2019b. Digitising Swiss German: How to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

T. Schmidt. 2016. Construction and dissemination of a corpus of spoken interaction: Tools and workflows in the FOLK project. In M. Kupietz and A. Geyken, editors, *Corpus Linguistic Software Tools*, pages 117–144.

T. Schmidt. 2018. Gesprächskorpora. In M. Kupietz and T. Schmidt, editors, *Korpuslinguistik*, pages 209–230. De Gruyter, Berlin and Boston.

T. Schmidt, W. Schütte, J. Winterscheid, M. Schürmann, S. Reineke, and E. Schedl. 2023. *cGAT: Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)*.

R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3):287–333.

M. Starosta. 1999. *Niedersorbisch-deutsches Wörterbuch*. Domowina-Verlag, Bautzen.

M. Toska. 2020. A rule-based normalization system for Greek noisy user-generated text. Master's thesis, Uppsala University.

L. Válková, M. Waclawičová, and M. Křen. 2012. Balanced data repository of spontaneous spoken Czech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 3345–3349, Istanbul. European Language Resources Association.

P. Völkel, editor. 2005. *Prawopisny słownik: Hornjoserbsko-němski słownik / Obersorbisch-deutsches Wörterbuch*, 5th edition. Domowina-Verlag, Bautzen.

S. Westpfahl and T. Schmidt. 2013. POS für(s) FOLK—Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. *Journal for Language Technology and Computational Linguistics*, 28(1):139–156.

H. Wiese, U. Freywald, S. Schalowski, and K. Mayr. 2012. Das KiezDeutsch-Korpus: Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache*, 40:97–123.

S. Yolchuyeva, G. Németh, and B. Gyires-Tóth. 2018. Text normalization with convolutional neural networks. *International Journal of Speech Technology*, 21(4).

C. Zimmer, H. Wiese, H. J. Simon, M. Zappen-Thomson, Y. Bracke, B. Stuhl, and T. Schmidt. 2020. Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations- und Soziolinguistik. *Deutsche Sprache*, 3:210–232.

12. Language Resource References

BNC Consortium. 2007. *The British National Corpus, XML Edition*. Oxford Text Archive.

Deppermann, A. and Hartung, M. and Schmidt, T. and Reineke, S. 2025. *Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*.

Frick, E. 2019. *Russlanddeutsche Dialektdatenbank (RuDiDat)*.

Földes, C. 2019. *Digitales Portal: Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus Projekt (UZSK)* .

Maitz, P. and König, W. 2022. *Unserdeutsch Rabaul Creole German (UNSD)*. Archiv für Gesprochenes Deutsch (AGD), IDS Mannheim.

Margaret Blevins. 2022. *Texas German Sample Corpus*.

Mueller, M. and Loewenstein, J. *Early Print Project*.

Texas German Dialect Project. 2025. *Texas German Dialect Project Corpus Platform*.

Sauer, S. 2016. *The Berlin Map Task Corpus (BeMaTaC)*.

Scherrer, Y. and Samardžić, T. and Glaser, E. 2019. *ArchiMob*.

Simon, H. and Wiese, H. 2019. *Deutsch in Namibia (DNam)*. Archiv für Gesprochenes Deutsch (AGD), IDS Mannheim.

Wiese, H. and Freywald, U. and Rehbein, I. and Schalowski, S. 2015. *Das KiezDeutsch-Korpus*.

Çetinoğlu, Ö. 2004. *SAGT: Computational Structural Analysis of German-Turkish Code-Switching*.

Handling Cross-Dialect Syntactic Variation: a Theory-Driven Web Resource

Emanuela Li Destri[1], Marco Longhin[1], Gaia Sorge[2], Sofia Ferroni[3], Giovanni B. Matteazzi[4], Andrea Artioli[2], Lorenzo Carletti[2], Federico Motta[2], Giuseppe Longobardi[5], Cristina Guardiano[*1,6]

[1]Dipartimento di Comunicazione ed Economia, Università di Modena e Reggio Emilia
[2]Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia
[3]Dipartimento di Studi Linguistici e Letterari, Università di Padova
[4]Dipartimento di Matematica, Università di Padova
[5]Department of Language and Linguistic Science, University of York
[6]Scuola Universitaria Superiore IUSS, Pavia
[*]cristina.guardiano@unimore.it

Abstract

Cross-dialect syntactic variation tests the limits of comparative analysis, owing to the entanglement of inheritance and contact in dialect systems. Addressing this challenge requires analytical tools combining the theoretical depth of formal models of grammatical competence with quantitative taxonomic techniques. The Parametric Comparison Method (PCM) embodies this integration by quantifying structural similarity across grammars through the comparison of abstract syntactic rules. The method has been shown to achieve a good degree of resolution in dialectal domains, capturing subtle contrasts and yielding configurations aligning with phylogenetic expectations while remaining sensitive to contact-induced convergence. Fully assessing its effectiveness as a resource for the quantitative study of syntactic dialectology, however, requires an infrastructure that ensures systematic data collection, consistent parameter setting, and dedicated computer-based taxonomic analyses. The PCM Hub is a web-based resource designed for this purpose. It integrates guided elicitation, automated parameter-setting procedures, data management, and the computation of distances and automatic classifications within a unified environment. By standardizing the transition from raw linguistic observations to a structured, replicable empirical apparatus, the PCM Hub provides the practical and quantitative support necessary to test the power of the PCM across expanded comparative domains.

Keywords: Web Resources, Data Collection and Analysis, Dialect Variation, Syntactic Comparison

1. Challenges for Dialect Comparison

Comparative approaches to cross-dialect syntactic diversity (Kayne, 1996, 2005; Poletto, 2012; Ledgeway, 2000; Ledgeway et al., 2018, 2020, a.o.) have demonstrated the relevance of investigating closely related systems to understand the minimal mechanisms underlying grammatical variation and reconstruct pathways of change by disentangling phylogenetic inheritance from interference.

In the framework of the Parametric Comparison Method (PCM; Longobardi and Guardiano, 2009), these issues converge in what is termed the 'ultralocality' problem (Guardiano et al., 2021, 146). The PCM, conceived for purposes of long-term historical reconstruction and phylogenetic inference, quantifies relations between languages through the comparison of syntactic *parameters* (Chomsky, 1981). The core assumption is that the distribution of parametric variation carries historical significance.

A first point of inquiry concerns the level of resolution achieved by parametric systems. The question is whether a system capable to encode syntactic variation at the *universal* level (Guardiano et al.'s 2021 *globality* problem) is also able to capture minimal differences among languages whose degree of relatedness stems from both 'vertical' (inheritance) and 'horizontal' (contact) historical paths. If so, a further question is whether the patterns of divergence identified correlate with a historically meaningful distribution.

These two dimensions have been addressed along two distinct lines of research. A first line investigates aspects of diversity between varieties that share a history of contact and are related to varying degrees of historical depth. Dedicated work has examined the extent to which the variation observed within Italo-Romance (Silvestri, 2013, 2016, 2018, 2020; Guardiano et al., 2022; Guardiano, 2023; Guardiano and Stalfieri, 2026, a.o.), or between Greek and Romance, particularly in Italiot Greek (Guardiano and Stavrou, 2014, 2019a,b, 2020, 2021), reflects vertical inheritance or horizontal diffusion, and how this information can be encoded in terms of parameter settings. A second line has focused on the mathematical modeling of the observed diversity by measuring parameter distances and assessing their distribution (Guardiano et al., 2016, 2021). This line of research has shown that "the PCM, even when applied to groups with minimal internal differentiation, is able to discrim-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 70–82
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

inate their articulations" (Guardiano et al., 2021, 172): parameters encode a signal that "is robustly tree-like, even in critical areas of patent language contact" (Guardiano et al., 2016, 152), while also capturing some degree of secondary convergence, whose quantitative assessment, however, still calls for methodological refinement.

2. Assessing Dialect Diversity Through the PCM

A radical point of departure of the PCM from attempts to lexically measure dialect diversity (Pellegrini, 1970; Goebl, 1982; Nerbonne and Kretzschmar, 2003; Nerbonne, 2009; Nerbonne and Heeringa, 2009) or even to use 'structural data' to infer language classifications lies in the nature and theoretical status of the linguistic material employed as taxonomic characters. In the PCM, the relevant taxonomic characters are parameter values rather than words, sounds, or observable structural patterns: languages are compared on the basis of the abstract *rules* that define their grammars rather than on the observable sequences of words or morphemes stemming from them (*manifestations*, Crisma et al., 2020).

In the PCM, the binary states of parameters are symbolized by [+] and [−]: [+] signals that the parameter is active in the language, i.e., that the language displays at least one of its manifestations; [−] signals that the parameter is absent from the grammar, i.e., none of its manifestations is observed.

Checking whether a parameter is active or not in a grammar (I-language, Chomsky, 1986) involves detecting, in the available data, the structures depending on that parameter: this can either be obtained from textual evidence (Crisma et al., 2025) or, as in classical formal practice, from grammaticality judgments of individual informants.

To this end, a parameter setting algorithm (Crisma et al., 2020) is required to identify the relevant triggering data, along with a practical tool to collect and systematically organize them.

Crucially, the parameter setting process does not require large quantities of data; it relies instead on carefully selected evidence. In this respect, a major challenge in dialectal settings is that speakers often find themselves in conditions of unbalanced bilingualism (Lambert et al., 1959), a situation common in endangered contexts (Polinsky, 1995; Grinevald, 2003; Bidese, 2017), which may make it difficult to assess the native status of their dialect competence and, consequently, the reliability of their judgments. Hence, assessing the speaker's competence becomes a particularly delicate task that must be handled with care (Cornips and Poletto, 2005; Bailey, 2017; Macaulay, 2017, a.o.). On these grounds, the internal architecture of the parametric framework offers a built-in control mechanism: data yielding parameter values inconsistent with the dependency structure of the system or with its theoretical predictions are immediately detectable, thus making it possible to isolate inaccurate information and neutralize its potential distortive effects.

The dependency structure is a distinctive feature of parameter systems (Baker, 2001; Fodor, 2001; Longobardi, 2003; Biberauer and Roberts, 2015, 2017; Roberts, 2019): one state of a given parameter may entail the irrelevance of another parameter, whose manifestations become predictable. Parameter dependencies are formalized as *implicational conditions*; neutralization due to a violated condition is encoded as [0] (Guardiano and Longobardi, 2017), which represents redundant information that should be excluded from comparison. A resource designed to manage a system of this kind must incorporate all of its defining properties and ensure that data collection and parameter setting are consistent with them. Such a tool should allow for linking data to the relevant parameter(s) and setting their values, while also taking into account the network of dependencies that may neutralize them.

The comparison of the resulting parameter lists requires additional tools to align them, measure distances and statistically assess their distribution, select subsets of languages and/or parameters to explore patterns of convergence and divergence in greater detail, derive classifications, and map the distribution of parametric distances and values across the geographical space.

The PCM Hub was developed to address these issues. This web-based application optimizes the conceptual architecture of the parameter system and the parameter setting algorithm by integrating them with a user-oriented data collection mechanism that meets accessibility standards, accommodates specific properties of different data types, and addresses the empirical and methodological requirements for automatic analyses. Designed to be flexible and expandable, this tool facilitates cumulative research, supports comprehensive parameter encoding and generates outputs suitable for statistical testing and phylogenetic reconstruction. Starting from a set of data supplied by a native speaker or language expert, each linked to the manifestation(s) of a given parameter, the system determines parameter values and evaluates the data against parameter dependencies. A resource that meets these requirements can, in principle, be used to collect, analyse, and process any type of linguistic data, including dialectal ones.

The following sections illustrate its architecture and core functionalities.

3. The PCM Hub

So far, cross-linguistic data collection for PCM investigations has been conducted in the absence of an infrastructure capable of systematically recording, managing and analysing linguistic material while incorporating the theoretical apparatus of the parameter system. The need to expand the comparative scope of the PCM in terms of typological range, genealogical depth, and empirical coverage has now made such an infrastructure indispensable. Recent efforts have therefore moved toward shareable environments, designed to uphold the PCM methodological standards while improving the efficiency of data acquisition, management, and sharing. A resource of this kind should ideally satisfy the following criteria:

1. Data must be recorded and shared through a user-oriented web-based interface.

2. A structured network of connections must be established among data, between data and their manifestations, between manifestations and parameters, and among parameters.

3. The consequences of parameter dependencies must be automatically computed in order to identify neutralized parameters.

4. The list of parameters and their properties must remain expandable and adaptable to revisions intrinsic to a framework that evolves alongside expanding empirical evidence and ongoing theoretical development.

5. Subsets of languages and/or parameters must be extractable, enabling systematic comparison and analysis.

6. Measures of dissimilarity and the resulting classifications must be automatically derived.

The scope of the PCM parameter system, as outlined above, is much wider than others in syntactic research, and has a deductive depth unparalleled in typology and quantitative phylogenetics. Consequently, no pre-existing software or modelling framework can be directly adapted to it.

The PCM Hub was conceived to be sufficiently rigorous to incorporate all structural aspects of the system, yet flexible enough to allow for integrations and revisions. Indeed, PCM research on ultralocality (Guardiano et al., 2016) has already shown that the analysis of microvariation may reveal aspects of variation not captured by the parameter set; this often requires the introduction of finer-grained options and a corresponding restructuring of the dataset (as shown in Fig. 1).

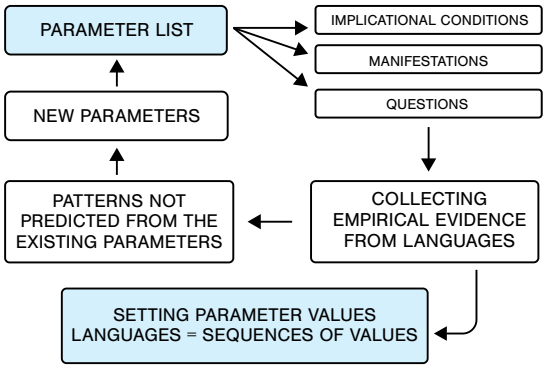


Figure 1: Parameter setting workflow

The infrastructure is inherently extensible: while the current parameter set reflects established theoretical knowledge, the platform is designed to seamlessly integrate new parameters as they emerge from the study of previously undocumented varieties or from further advancements in syntactic theory.

The platform has been designed with a view to future integration of phylogenetic inference, and will provide a usable and accessible tool to implement PCM analyses, while ensuring transparency of all methodological steps, in adherence to the FAIR principles (Wilkinson et al., 2016). In this sense, it is conceived primarily as a dedicated resource for PCM investigation, enabling researchers to carry out the workflow even in the absence of highly specialized expertise, while not positioning itself as a competitor to existing large-scale databases. Still, such resources, especially when grounded in rigorous principles of data collection and syntactic analysis – such as, in the Italo-Romance domain, the recent web platform based on Manzini and Savoia (2005) – may constitute a valuable source of empirical evidence. Interoperability with other databases, in line with standardization initiatives such as Forkel et al.'s (2018) Cross-Linguistic Data Formats (CLDF), is therefore envisaged as a potential feature of the infrastructure, although it is unlikely to provide a decisive solution for the full automation of data collection.

3.1. Interface overview

In its current version, the PCM Hub supports three types of activities: (a) data collection; (b) data analysis, including parameter setting, extraction of parameter sequences for statistical and phylogenetic processing, crosslinguistic and cross-parametric searches; and (c) retrieval of language information and procedural workflow. These activities correspond to three access modalities: (a) *User* (for informants, data collectors, language specialists),

(b) *Admin* (for PCM researchers), and (c) *Public Access* (for research replication). So far, development has primarily focused on the User and Admin interfaces, discussed below. Regarding Public Access, links to published PCM work (such as lists of languages with technical details and examples for each parameter, maps, dendrograms, comparative graphs and taxonomies) will be made accessible.

3.1.1. Data collection: the *User* interface

The data-entry platform is designed to allow the autonomous collection and entry of targeted evidence associated with the manifestations of each parameter by trained native speakers (or language experts). Since language experts may lack expertise in syntactic analysis, elicitation must be carefully guided (Koopman and Guardiano, 2022). To this end, a step-by-step procedure that presents the relevant syntactic environments needed to test the target structures was designed following the data collection protocol presented in Crisma et al. (2020) and Crisma et al. (2025). This procedure does not rely on standardized questionnaires. Instead, it offers a description of the structure(s) required to set the parameter, together with one or more examples from languages in which the relevant configuration is unambiguously attested. Doing so ensures that technical details are correctly interpreted by contributors. Language experts are then asked to provide their own examples, thus being free to make their own lexical choices and adapt the examples to language-specific needs. The system's support for IPA transcription makes the platform fully operational even for languages lacking a written standard. Finally, it is important to note that, although the User interface has been designed to make data entry as straightforward as possible, the resource is intended for expert use and does not accommodate untrained speakers, whose input therefore requires expert mediation and guidance.

The workflow is structured as follows. For each parameter, the interface presents a concise description alongside a series of YES/NO questions corresponding to its manifestations. After a positive answer is given, the User is required to submit a minimum of two examples, along with their transliteration (where necessary), gloss, and translation (see Figure 2). Conversely, selecting a negative answer requires the User to choose one of the predefined explanations accounting for that response (see Figure 3).

Each stage is supported by specific instructions, a glossary of technical terminology, and concise glossing guidelines. A dedicated Comment field allows for further observations/comments. This feature enables researchers to document gradient phenomena or cases that might, at first glance, be difficult to fit into a strict binary classification, ensuring that the quantitative abstraction is always supported by nuanced linguistic evidence.

Data reliability is monitored through a dual-color confidence system: a green button for complete and verified data, a red one for further review (Figure 4). Until final submission, entries remain editable. Users may download the data they contribute, which will be credited to them in relevant PCM publications.

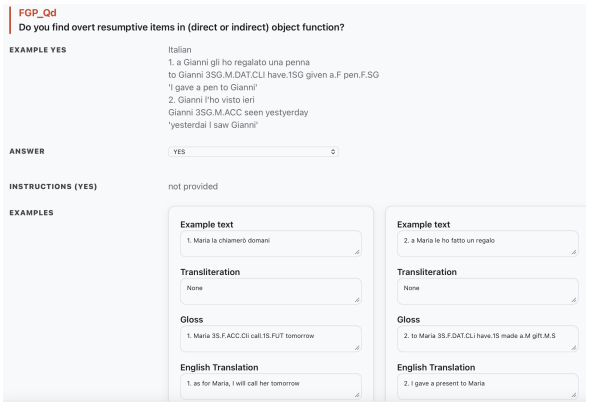


Figure 2: Data entry interface, answer YES

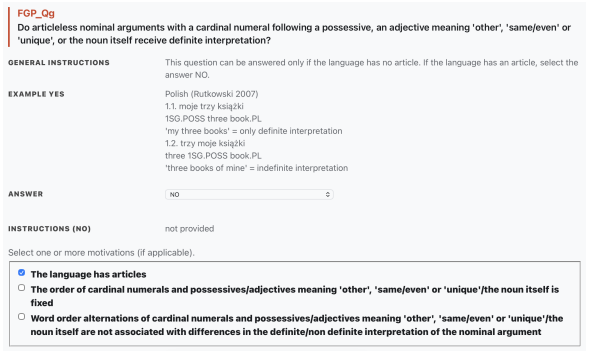


Figure 3: Data entry interface, answer NO



Figure 4: Data entry interface, confidence system

3.1.2. Data analysis: the *Admin* interface

Admins oversee access to the dataset and control data entry. They assign each language expert (User) to a specific language and enter relevant metadata, including standard identification codes (Glottocode and ISO code), genealogical classification (when available), and geographical coordi-

nates. These data can be exported and used for different types of analysis (Section 4.3).

Admins review the data to ensure that examples are consistent with the answer provided. If inconsistencies are detected, Users may be asked to modify or refine their entries. This process ensures quality control and uniformity.

The data collected as part of previous PCM publications are uploaded via structured files, which are automatically integrated into the platform.

Parameter values are computed through an algorithm that processes the answers in combination with the system of conditions formalizing cross-parametric dependencies. At the end of this process, each language is expressed as a sequence of symbols ([+], [−], and [0]).

Admins oversee database content: they can add, edit or deactivate parameters, link each parameter to its corresponding schema, type, set of manifestations, examples, and implicational conditions (Figure 5); edit the support material for Users. Parameters can be deactivated without loss of associated data, as the system ensures full preservation and traceability of all materials entered over time.

The platform supports cross-linguistic comparison by allowing Admins to select subsets of languages and/or parameters and to compute pairwise distances using dedicated algorithms. It enables multivariate techniques such as Principal Component Analysis (PCA) and hierarchical clustering visualized as dendrograms; it includes geographic visualization tools for generating maps such as those illustrated in Figure 8.

Finally, the system provides dedicated query tools for specific analysis of the distribution of parameter values and their manifestations.

All materials can be downloaded in multiple formats, including non-editable formats for dissemination (PDF) and machine-readable formats suitable for further processing through external scripts and computer-based tools not natively supported by the platform.

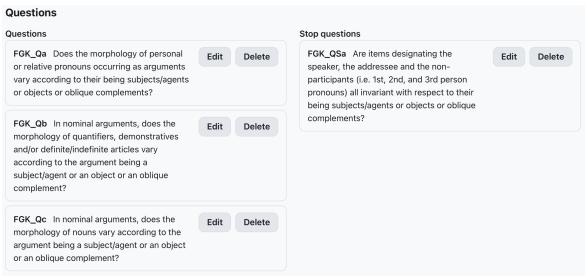


Figure 5: Question editing interface

3.2. Technical Specs

The PCM Hub is a scalable platform developed using the Django framework and a PostgreSQL backend. Its architecture follows a clear separation between data management, application logic, and user interface, ensuring stability, maintainability, and future extensibility. This approach allows the system to consistently handle the progressive increase in the number of languages, parameters, and operators. The interface is developed with responsive web technologies (HTML5, CSS, JavaScript). Each language and parameter is represented as a relational entity in the database, allowing for a structured and queryable organization of data. Parameter states are automatically assigned in response to the User's answers. Subsequently, a dedicated logical engine currently based on directed acyclic graphs (DAGs) assigns a [0] to those parameters whose implicational condition is not satisfied.

All software and technologies employed are open source; by contrast, the platform design, its structure and all of its content are protected by copyright.

4. Dialect Analysis and Comparison

This section illustrates some key functionalities of the PCM Hub in their application to ultralocality.

4.1. Languages

The current PCM dialect dataset consists of a selection of Romance and Greek dialects. The Romance dialects are all spoken in Italy, and are representative of six independently recognized groups in this domain, spanning across a north-south geographical space: (1) Ladin (Rocca Pietore); (2) Venetian (Trieste, Motta di Livenza, Vicenza, Chioggia, Porto Viro); (3) Northern Italy Gallo-Italic (Casalmaggiore, Parma, Reggio Emilia, Novellara, Correggio, Savignano sul Rubicone); (4) Upper Southern (Teramo, Santa Maria Capua Vetere, Frattamaggiore, Palma Campania, Felitto, Francavilla in Sinni, Verbicaro, Gargano, Barletta, Bari, Taranto); (5) Extreme Southern (Mesagne, Cellino San Marco, Botrugno, Cutro, Nicastro, Catanzaro, Reggio Calabria, San Filippo del Mela, Trapani, Ribera, Sant'Angelo Muxaro, Mussomeli, Catania, Ragusa); (6) Gallo-Italic of Sicily (Nicosia, Aidone). The Greek dialects instantiate three major groups, ideally representing the major compartments of the Greek-speaking world: (1) Italiot Greek (Grecìa Salentina, Bovesìa – two varieties); (2) Asia Minor Greek (Romeyka Pontic – Sitaridou, 2014; Schreiber, 2024; Neocleous and Sitaridou, 2025, Cappadocian – Karatsareas, 2009, 2013, 2016, Pharasiot – Bağrıaçık, 2017, 2018; Bağrıaçık and Danckaert, 2022); (3) 'Central' Greek (Standard Modern Greek, Cypriot).

This selection allows for testing ultralocality questions beyond those addressed in previous PCM studies. Once the tree-like signal identified in earlier work (namely, the distinction of the Romance cluster from Greek) is replicated, the next question is whether finer-grained internal structure can be detected within each cluster. For example, (i) Does the distribution of parametric diversity align with traditionally recognized classifications of the Romance dialects of Italy? (ii) Does it recover the tripartite structure of the Greek-speaking domain?

The data required to set each parameter were collected over the period 2010-2025, before the PCM Hub was implemented. These data are being stored in the PCM Hub through structured files, following the procedure described in Section 3.1.2.

As observed in Section 2, the lists of parameter values used for comparison are assumed to instantiate the competence of individual speakers, such that, in principle, each speaker is associated with a distinct list of values. Ideally, each list should therefore be constructed consistently on the basis of the judgments of a single individual speaker. In principle, judgments provided by speakers of the same language/dialect are expected to largely overlap. Accordingly, when judgments from different speakers of the same language/dialect converge, only one list of values is produced. By contrast, when different speakers provide judgments which lead to opposite settings for one or more parameters, multiple lists of values are produced, one for each speaker.

This idealized procedure, however, cannot always be implemented with full precision, as in practice, circumstances may arise in which it is not possible to obtain all the required judgments from a single speaker (for instance, when updates to the parameter system make it necessary to elicit additional judgments and the original speaker is no longer available). In such cases, corrective measures are adopted, including consultation of the relevant literature.

In the present dataset, for some dialects (including Gargano, Nicosia, Asia Minor Greek, and one variety of Italiot Greek from Bovesìa), due either to the unavailability of native speakers or to the impossibility of maintaining systematic contact with the same speaker, data were collected from written sources or from the existing literature, in accordance (when possible) with the protocol defined in Crisma et al. (2025).

The parameter dataset currently implemented in the platform consists of the 94 parameters described in Crisma et al. (2025), with the addition of 14 further points of variation tentatively formulated to account for aspects of microparametric diversity detected across the Romance sample and briefly presented in Guardiano and Stalfieri (2026).

The lists of parameter values corresponding to each dialect are analysed and compared using exploratory statistical analyses, clustering techniques, and mapping algorithms implemented within the platform through dedicated scripts. Concerning statistical and computer-assisted analyses, a range of approaches have been applied over the years to various PCM datasets (Bortolussi et al., 2011; Guardiano et al., 2016; Franzoi et al., 2019; Ceolin et al., 2021; Guardiano and Stalfieri, 2026; see Longhin et al. (in prep.) for a detailed methodology-based comparative discussion). These studies consistently show that the results are robust across methods, with minimal variation in the resulting taxonomic classifications. The three methods presented in what follows are intended to provide the clearest illustrative examples.

4.2. Principal Component Analysis

The Principal Component Analysis (PCA; Jolliffe, 2002; Jolliffe and Cadima, 2016; Gewers et al., 2021) is an exploratory statistical technique that extracts orthogonal components from a matrix of variables. Despite distortions introduced by binary encoding of parameter states, the PCA provides more interpretable representations than other multivariate techniques and has yielded the most transparent results across different language datasets. An additional advantage is that it enables the identification of the variables shaping the observed parametric diversity. However, more refined analytical approaches may prove even more effective in future applications.

The analysis begins by removing those variables (i.e., parameters) that behave as constants, i.e., whose values are identical across all taxonomic units (the dialects). On this basis, 61 out of 108 variables were excluded. The distribution of parameter values across the taxonomic units can therefore be analysed only on a reduced set of variables, as is to be expected in a dataset of genealogically closely related languages (Guardiano and Longobardi, 2005). Despite this reduction, the two principal components account for 38.69% of the total variance, and the distribution of the taxonomic units across the space defined by these two components (F1, 25.89%, and F2, 12.80%), as illustrated in the scatter plot in Figure 6, remains highly informative.

This distribution can first be interpreted based on F1 scores. The Greek varieties are associated with positive F1 scores, as opposed to the Romance dialects, all displaying negative F1 scores. This result provides empirical corroboration of the expectation that parameter values identify a Greek cluster separate from Romance. This supports the conclusion that, although effects of contact-induced parameter resettings are visible in the two domains (see Section 1), inherited patterns in parameter transmission

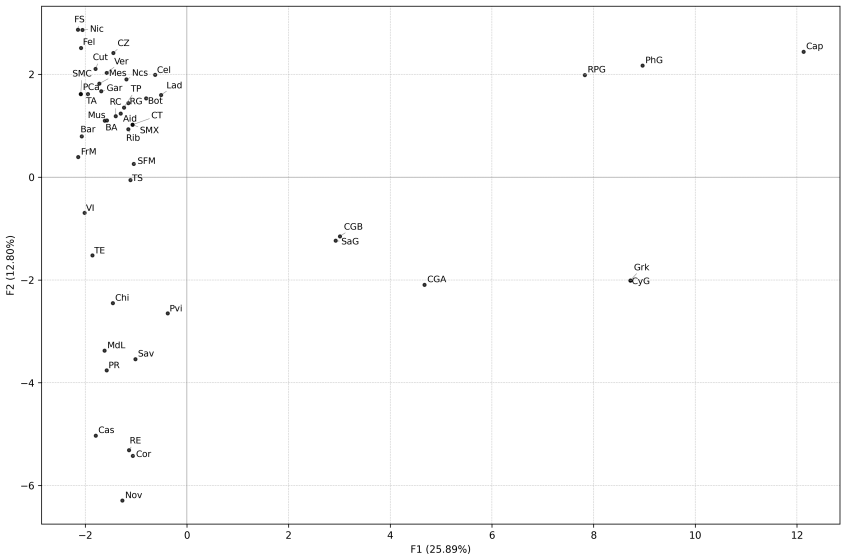


Figure 6: Scatter plot of the PCA

are not obscured by horizontal convergence.

It must also be noted that, while the Romance dialects are compressed within a narrow range of negative F1 scores, the Greek varieties are distributed across a much broader area, suggesting a sharper internal dialect differentiation.

The combination of F1 and F2 scores defines three Greek clusters, corresponding to Asia Minor Greek (highly positive F1 and F2 scores), Central Greek (highly positive F1 scores, moderately negative F2 scores), and Italiot Greek (moderately positive F1 scores, moderately negative F2 scores). This provides a positive answer to question (ii): the tripartite structure of the Greek-speaking domain is reflected in the distribution of parameter values.

The distribution of the Romance dialects is considerably less clear-cut, although some groupings can still be identified. Negative F2 scores separate a central-northern geographical area (except for Ladin) from southern dialects (except for Teramo), all of which display positive F2 scores. In this quadrant, no further clearly identifiable subdivisions emerge. By contrast, in the quadrant corresponding to negative F2 scores, the dialects are more dispersed. The most notable exception to the North-South divide captured by F2 scores is Ladin, a Rhaeto-Romance dialect group spoken in Northern Italy, whose classification and internal relations have been the subject of long-standing debate (Ascoli, 1873; Battisti, 1931; Haiman and Benincà, 2005). In the absence of closely related dialects such as Romansh or Friulian, Ladin constitutes an isolate in the sample. Hence, no strong a priori prediction can be made regarding its placement. The fact that it does not display any marked similarity with the other dialects of Northern Italy indicates that the distribution of parametric diversity cannot be reduced to a mere function of geographical proximity.

4.3. Hierarchical cluster analysis

Hierarchical agglomerative clustering methods (Everitt et al., 2011) take dissimilarity matrices as input and implement several linkage criteria, including: (1) Single linkage (minimum inter-cluster distance: nearest neighbor – Johnson, 1967; Sneath and Sokal, 1973); (2) Complete linkage (maximum inter-cluster distance: farthest neighbor – Johnson, 1967; Sneath and Sokal, 1973); (3) Average linkage (mean inter-cluster distance: UPGMA – Sokal and Michener, 1958; Lance and Williams, 1967); (4) Ward's method, which relies on Euclidean distances and is applied to the character matrix to minimize the increase in within-cluster variance at each agglomerative step (Ward, 1963).

These algorithms are implemented in the PCM Hub. Although comparable procedures are available in other web-based platforms for dialect data analysis (e.g., Gabmap, Nerbonne et al., 2011), their integration within the PCM Hub enables analyses to be conducted within a continuous workflow, eliminating the need to adapt PCM data formats to external platforms. This minimizes the risk of information loss. It also allows PCM researchers to operate on a set of algorithms directly suited to the analysis of parametric datasets.

To obtain a dissimilarity matrix from the lists of parameter values, pairwise parametric distances are computed using a formula traditionally adopted in PCM works (Ceolin et al., 2020), here simplified as *d/(i+d)*, where *d* denotes the number of differences in parameter states and *i* the number of identities. The latter can be calculated either by counting all

cases in which both languages display the same value ([+] or [−]) for a given parameter, or by counting only the cases in which both display [+]. The analyses presented here adopt the former criterion, although, at this level, both yield virtually identical results. Pairs in which one or both languages display a [0] are excluded from computations.

Figure 7 presents a dendrogram generated using the average linking method (UPGMA). The dendrogram reflects a taxonomic structure consistent with the PCA results: a clear primary split into two nodes corresponding to Greek and Romance, followed by an internal articulation of each cluster that largely mirrors the established internal structure of the groups. Within the Greek cluster, the outermost nodes separate Asia Minor Greek from a group including the central varieties and Italiot Greek. Within the latter, the two currently spoken varieties form a distinct cluster, as opposed to the conservative variety of Bovesìa (CGA), which is closer to Central Greek. This result is consistent with proposals suggesting that contact-induced effects from Romance represent relatively recent innovations (Guardiano and Stavrou, 2021). The articulation of Romance is largely consistent with independently established dialect classifications, though with some noteworthy divergences. Leaving aside the position of Ladin, already discussed in Section 4.2, two results are of particular interest. First, the Gallo-Italic dialects of Sicily are located within the cluster grouping the Extreme Southern varieties. This outcome is in line with recent work on the syntax of the Gallo-Italic dialects of Sicily (De Angelis, 2023; Guardiano and Stalfieri, 2026, a.o.). Second, the dialects spoken in central Calabria cluster with the Upper dialects, rather than with the Extreme Southern group. This configuration aligns with the literature identifying the corresponding area as a transitional zone of uncertain classification (Pellegrini, 1977; Trumper and Maddalon, 1988; Trumper, 1997, a.o.). Overall, these results provide a positive answer to question (i) in Section 4.1, as the only inconsistencies with traditional dialect classifications concern already controversial areas.

The clustering structure and its robustness were further tested using a second algorithm, K-means clustering (Forgy, 1965; Macqueen, 1967; Lloyd, 1982), which partitions taxonomic units on the basis of their parameter-value vectors by minimizing within-cluster variance (typically computed using Euclidean distance). These analyses yield a taxonomic structure closely comparable to that observed in Figure 7.

A dedicated script was developed to project the clustering results onto a geographical map. The script takes as input a table of geographical coordinates for each language, automatically generated within the PCM Hub. Once a predefined number of clusters is selected, the algorithm produces a map assigning each language to one of the clusters identified in the tree. Figure 8 was generated by setting the number of clusters to two. Increasing the number of clusters yields progressively finer partitions, which, as noted above, largely correspond to traditionally established dialect classification.

Algorithms for the areal visualization of cluster structures are standard in dialectometry (Goebl, 1982; Nerbonne and Kretzschmar, 2003; Nerbonne, 2009; Nerbonne and Heeringa, 2009, see also Dialektkarten.ch) and are implemented in web-based platforms such as Gabmap. These procedures model the spatial distribution of features, or feature bundles, across a territorially discretized space, represented through granular polygonal tessellations. While these tools have been shown to perform well even on sparse datasets, their resolution improves as geographical coverage becomes denser, since they rely on the cartographic projection of feature-based distance measures across finely articulated spatial grids. At present, the high degree of spatial granularity required for these techniques to be truly informative is as yet only partially available in the PCM. Additionally, this type of cartographic representation is methodologically better suited to examining the geographical diffusion of individual phenomena than to extracting historical signals from the distribution of multiple features. These are therefore more appropriate for mapping the distribution of individual parameter values, provided sufficiently fine-grained geographical coverage. Currently, PCM analyses use cartographic visualization to assess the extent to which parametric similarity is predicted by geographical proximity. For this purpose, a lightweight algorithm that directly projects cluster assignments onto a geographical map using information already stored within the PCM Hub offers a more efficient solution. At the same time, the PCM Hub does not position itself as an alternative to existing platforms. Owing to its data-export functionalities in multiple formats, it allows PCM datasets to be used in a wide range of external analytical environments. It should therefore be understood not as a replacement, but as a complementary theory-oriented infrastructure for quantitative dialect analysis.

Further evidence challenging the possibility that parametric similarity can be entirely reduced to geographical proximity is provided by Mantel tests (Mantel, 1967; Sokal and Rohlf, 1995; Guillot and François, 2013) correlating parametric and geographical distance matrices. The correlation coefficient (Pearson's r=0.5904, p=0.0010; r=0.5157, p=0.0010 when limited to the Italian peninsula; r=0.59, p=0.0010 if restricted to Romance) indicates a statistically significant and relatively strong

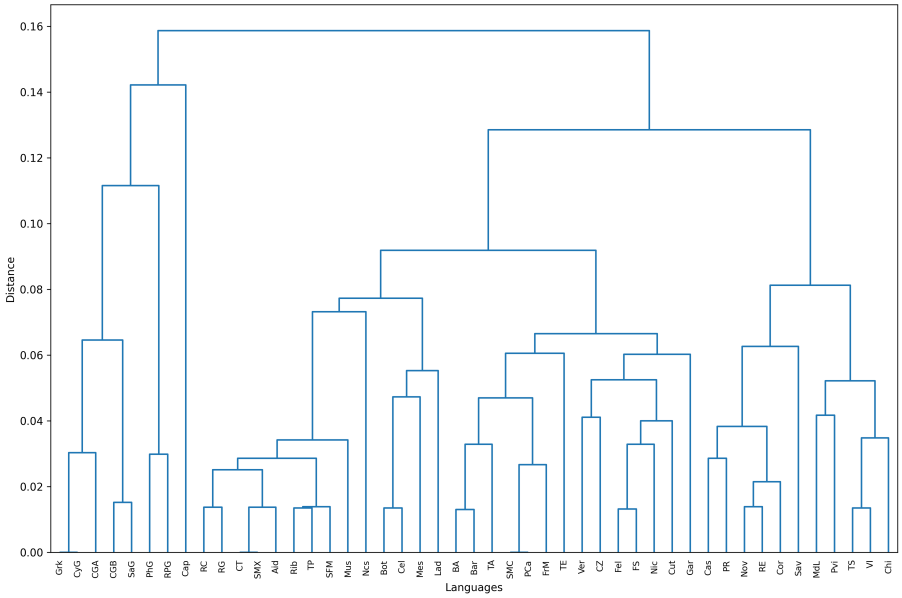


Figure 7: Hierarchical clustering dendrogram

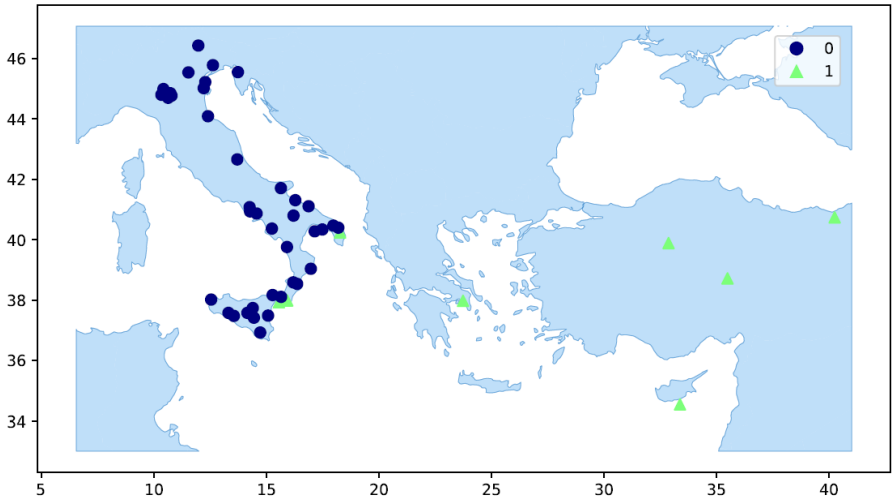


Figure 8: K-means clustering projected onto a map

association, yet not one sufficient to support direct predictability. Even within a geographically restricted and contact-intensive domain, geographical distance does not fully account for linguistic proximity, suggesting that the signal encoded in syntactic parameters remains robust enough to preserve genealogical differentiation despite areal interference.

5. Conclusion

The PCM approach to ultralocality demonstrates that dialect syntactic diversity can be measured in a formally constrained, reproducible and meaningful way. It introduces a quantitative perspective on ultralocal syntactic variation that, while building on dialectometric traditions, geolinguistic models, and formal treatments of syntactic microvariation, extends beyond them. The applications illustrated in this contribution suggest that the transmission of syntactic parameters, while sensitive to areal convergence, preserves a clear inheritance structure even in ultralocal domains. A more granular investigation of vertical transmission and horizontal diffusion in these domains is therefore poised to contribute to an integrated model of local variation, while also identifying patterns of syntactic transmission applicable to a deeper historical scale. In this light, the PCM Hub emerges as a key resource ensuring a theory-oriented empirical infrastructure to explore the parametric diversity of dialects.

6. Acknowledgements

Part of this research was funded by the Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN), project "Measuring the power of parameter setting theory on historical corpora – PARTHICO", Grant Assignment Decree n. 901, 21/06/2023, Italian Ministry of University and Research – MUR, and by the Progetto di Ricerca Interdisciplinare FAR Unimore 2024 Modeling crosslinguistic diversity in Differential Object Marking through the Parametric Comparison Method (POM; prot. n. 0323715 del 28/11/2024, Rep. n. 1307/2024).

Most of the data from the Romance dialects used in this work were collected by Vincenzo Stalfieri as part of the project n. Prin2017-2017K3NHHY, Models of language variation and change: new evidence from language contact, Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN), and analysed with the advice of Franco Fanciullo.

The data from the Greek domain were collected and analysed as part of collaborative work with Melita Stavrou.

The structure, content and design of the web platform were discussed many times with Paola Crisma, who provided essential advice.

Access to the list of informants, geographical coordinates and data for each dialect will be accessible through the website www.parametricomparison.unimore.it, where the link to the PCM Hub, along with detailed instructions to access the platform, will be made available.

All scripts and technical tools used to generate the results presented in this study are publicly available at https://github.com/GaiaSorge/PCM_Hub.

Author contribution. Conceptualization and methodology: CG and GL. Software development: ML and GBM. Data collection and analysis: SF, ELD, ML, CG, GS. Scripts and computer-assisted analyses: AA, ML, LC and FM. Writing - first draft: ELD, GS and ML. Writing - review and editing: CG, SF and GL (all authors, final version). Supervision, Funding acquisition, Project administration: CG.

7. References

Graziadio Isaia Ascoli. 1873. Saggi ladini. In Graziadio Isaia Ascoli, editor, *Archivio glottologico italiano*, volume 1, pages 1–537. Loescher, Roma/Torino/Firenze.

Guy Bailey. 2017. Field interviews in dialectology. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 284–299. Wiley, Hoboken.

Mark C. Baker. 2001. *The atoms of language. The mind's hidden rules of grammar*. Oxford University Press, Oxford.

Carlo Battisti. 1931. *Popoli e lingue nell'Alto Adige. Studi sulla latinità altoatesina*. Bemporad, Firenze.

Metin Bağrıaçık. 2017. Representing discourse in clausal syntax. The *ki* particle in Pharasiot Greek. *Journal of Greek Linguistics*, 17:141–189.

Metin Bağrıaçık. 2018. *Pharasiot Greek: word order and clause structure*. Ph.D. thesis, Ghent University, Ghent.

Metin Bağrıaçık and Lieven Danckaert. 2022. Raising and matching in Pharasiot Greek relative clauses. A diachronic reconstruction. *Journal of Linguistics*, 58(3):495–533.

Theresa Biberauer and Ian Roberts. 2015. The clausal hierarchy, features and parameters. In Ur Shlonsky, editor, *Beyond Functional Sequence*, pages 295–313. Oxford University Press, Oxford.

Theresa Biberauer and Ian Roberts. 2017. Parameter setting. In Adam Ledgeway and Ian Roberts, editors, *The Cambridge Handbook of Historical Syntax*, pages 134–162. Cambridge University Press, Cambridge.

Ermenegildo Bidese. 2017. The correlation between unbalanced bilingualism and language decay in small language minorities. The current status of research and future perspectives. *Bollettino dell'Atlante linguistico italiano*, 41:95–107.

Luca Bortolussi, Andrea Sgarro, Giuseppe Longobardi, and Cristina Guardiano. 2011. How many possible languages are there? In Gemma Bel-Enguix, Veronica Dahl, and M. Dolores Jiménez-López, editors, *Biology, computation and linguistics*, pages 168–179. IOS Press, Amsterdam.

Andrea Ceolin, Cristina Guardiano, Giuseppe Longobardi, and Monica Alexandrina Irimia. 2020. Formal syntax and deep history. *Frontiers in psychology*, 11.

Andrea Ceolin, Cristina Guardiano, Giuseppe Longobardi, Monica Alexandrina Irimia, Luca Bortolussi, and Andrea Sgarro. 2021. At the boundaries of syntactic prehistory. *Philosophical Transactions of the Royal Society B*, 376:20200197.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

Noam Chomsky. 1986. *Knowledge of language. Its nature, origin, and use*. Praeger, New York.

Leonie Cornips and Cecilia Poletto. 2005. On standardising syntactic elicitation techniques (part 1). *Lingua*, 115(7):939–957.

Paola Crisma, Giulia Fabbris, Giuseppe Longobardi, and Cristina Guardiano. 2025. What are your values? Default and asymmetry in parameter states. *Journal of Historical Syntax*, 9:1–26.

Paola Crisma, Cristina Guardiano, and Giuseppe Longobardi. 2020. Syntactic diversity and language learnability. *Studi e Saggi Linguistici*, 58(2):99–130.

Alessandro De Angelis. 2023. The strange case of the Gallo-Italic dialects of Sicily. Preservation and innovation in contact-induced change. *Languages*, 8(3):163.

Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster analysis*. John Wiley and Sons, Chichester.

Janet Dean Fodor. 2001. Setting syntactic parameters. In Mark Baltin and Chris Collins, editors, *The Handbook of Contemporary Syntactic Theory*, pages 730–767. Blackwell, Oxford.

Edward W. Forgy. 1965. Cluster analysis of multivariate data. Efficiency versus interpretability of classifications. *Biometrics*, 21:768–780.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205.

Laura Franzoi, Andrea Sgarro, Anca Dinu, and Liviu P. Dinu. 2019. Linguistic classification: dealing jointly with irrelevance and inconsistency. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 345–352, Varna. INCOMA.

Felipe L. Gewers, Gustavo R. Ferreira, Henrique F. De Arruda, Filipi N. Silva, Cesar H. Comin, Diego R. Amancio, and Luciano Da F. Costa. 2021. Principal component analysis. A natural approach to data exploration. *ACM Computing Surveys*, 54(4):1–34.

Hans Goebl. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

Colette Grinevald. 2003. Speakers and documentation of endangered languages. *Language documentation and description*, 1:52–72.

Cristina Guardiano. 2023. Differential object marking in a dialect of Sicily. In Monica Alexandrina Irimia and Alexandru Mardale, editors, *Differential Object Marking in Romance. Towards microvariation*, pages 192–231. John Benjamins, Amsterdam.

Cristina Guardiano, Michela Cambria, and Vincenzo Stalfieri. 2022. Number morphology and bare nouns in some Romance dialects of Italy. *Languages*, 7(4):255.

Cristina Guardiano and Giuseppe Longobardi. 2005. Parametric comparison and language taxonomy. In Montserrat Batllori, Maria-Lluïsa Hernanz, Carme Picallo, and Francesc Roca, editors, *Grammaticalization and Parametric Variation*, pages 149–174. Oxford University Press, Oxford.

Cristina Guardiano and Giuseppe Longobardi. 2017. Parameter theory and parametric comparison. In Ian Roberts, editor, *The Oxford Handbook of Universal Grammar*, pages 377–398. Oxford University Press, Oxford.

Cristina Guardiano, Giuseppe Longobardi, Guido Cordoni, and Paola Crisma. 2021. Formal syntax as a phylogenetic method. In Richard D. Janda, Brian D. Joseph, and Barbara S. Vance, editors, *The Handbook of Historical Linguistics*, volume 2, pages 145–182. Wiley-Blackwell, Malden.

Cristina Guardiano, Dimitris Michelioudakis, Andrea Ceolin, Monica Irimia, Giuseppe Longobardi, Nina Radkevich, Giuseppina Silvestri, and Ioanna Sitaridou. 2016. South by Southeast. A syntactic approach to Greek and Romance microvariation. *L'Italia dialettale*, LXXVII:95–166.

Cristina Guardiano and Vincenzo Stalfieri. 2026. Comparazione parametrica e strutture nominali. Alcune note dalla Sicilia galloitalica. In Elvira Assenza, Angela Castiglione, Alessandro De Angelis, and Salvatore Menza, editors, *I dialetti galloitalici di Sicilia tra resistenza e assimilazione*, pages 163–192. Centro Studi Filologici Linguistici Siciliani, Palermo.

Cristina Guardiano and Melita Stavrou. 2014. Greek and Romance in Southern Italy. History and contact in nominal structures. *L'Italia dialettale*, LXXV:121–147.

Cristina Guardiano and Melita Stavrou. 2019a. Adjective-noun combinations in Romance and Greek of Southern Italy. Polydefiniteness revisited. *Journal of Greek Linguistics*, 19:3–57.

Cristina Guardiano and Melita Stavrou. 2019b. Comparing patterns of adjectival modification in Greek. A diachronic approach. *Quaderni di Linguistica e Studi Orientali*, 5:135–173.

Cristina Guardiano and Melita Stavrou. 2020. Dialectal syntax between persistence and change. The case of Greek demonstratives. *L'Italia dialettale*, 81:119–155.

Cristina Guardiano and Melita Stavrou. 2021. Modeling syntactic change under contact. The case of Italiot Greek. *Languages*, 6(2):74.

Gilles Guillot and Rousset François. 2013. Dismantling the Mantel tests. *Methods in ecology and evolution*, 4:336–344.

John Haiman and Paola Benincà, editors. 2005. *The Rhaeto-Romance languages*. Routledge, London/New York.

Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Ian T. Jolliffe. 2002. *Principal Component Analysis*. Springer-Verlag, New York.

Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis. A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374:20150202.

Petros Karatsareas. 2009. The loss of grammatical gender in Cappadocian Greek. *Transactions of the Philological Society*, 107(2):196–230.

Petros Karatsareas. 2013. Understanding diachronic change in Cappadocian Greek. The dialectological perspective. *Journal of Historical Linguistics*, 3(2):192–229.

Petros Karatsareas. 2016. Convergence in word structure. Revisiting agglutinative noun inflection in Cappadocian Greek. *Diachronica*, 33(1):31–66.

Richard S. Kayne. 1996. Microparametric syntax. Some introductory remarks. In James R. Black and Virginia Motapanyane, editors, *Microparametric Syntax and Dialect Variation*, volume 139, pages ix–xviii. John Benjamins, Amsterdam.

Richard S. Kayne. 2005. Some notes on comparative syntax. With special reference to English and French. In *Movement and Silence*, pages 277–334. Oxford University Press, Oxford.

Hilda Koopman and Cristina Guardiano. 2022. Managing data in TerraLing, a large-scale cross-linguistic database of morphological, syntactic, and semantic patterns. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, editors, *The Open Handbook of Linguistic Data Management*, pages 617–630. The MIT Press, Cambridge, MA.

Wallace E. Lambert, Jelena Havelka, and Richard C. Gardner. 1959. Linguistic manifestations of bilingualism. *The American Journal of Psychology*, 72(1):77–82.

Godfrey N. Lance and William T. Williams. 1967. A general theory of classificatory sorting strategies. 1. Hierarchical systems. *The Computer Journal*, 9(4):373–380.

Adam Ledgeway. 2000. *A comparative syntax of the dialects of southern Italy. A minimalist approach*. Blackwell, Oxford.

Adam Ledgeway, Norma Schifano, and Giuseppina Silvestri. 2018. Il contatto tra il greco e le varietà romanze nella Calabria meridionale. *Lingue antiche e moderne*, 7:96–133.

Adam Ledgeway, Norma Schifano, and Giuseppina Silvestri. 2020. Microvariation in dative-marking in the Romance and Greek varieties of Southern Italy. In Anna Pineda and Jaume Mateu, editors, *Dative constructions in Romance and beyond*, pages 311–342. Open Generative Syntax, Language Science Press.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Marco Longhin, Isabella Morlini, Gaia Sorge, Monica A. Irimia, and Cristina Guardiano. in prep. Parametric comparison and phylogenetic testing. Ms. UniMoRe.

Giuseppe Longobardi. 2003. Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook*, 3(1):101–138.

Giuseppe Longobardi and Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119:1679–1706.

Ronald Macaulay. 2017. Dialect sampling methods. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 241–252. Wiley, Hoboken.

James Macqueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkley. University of California Press.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2):209–220.

Maria Rita Manzini and Leonardo Maria Savoia. 2005. *I dialetti italiani e romanci. Morfosintassi generativa*. Edizioni dell'Orso, Alessandria.

Nicolaos Neocleous and Ioanna Sitaridou. 2025. Word order and information structure in Romeyka. *Frontiers in Psychology*, 16:1337962.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap – a web application for dialectology. *Dialectologia*, Special issue II:65–89.

John Nerbonne and Wilbert Heeringa. 2009. Measuring dialect differences. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space: Theories and Methods*, pages 550–567. De Gruyter Mouton, Berlin/New York.

John Nerbonne and William Kretzschmar. 2003. Introducing computational techniques in dialectometry. *Computers and the Humanities*, 37:245–255.

Giovan Battista Pellegrini. 1970. La classificazione delle lingue romanze e i dialetti italiani. *Forum Italicum: A Journal of Italian Studies*, 4(2):211–237.

Giovan Battista Pellegrini. 1977. *Carta dei dialetti d'Italia*. Pacini, Pisa.

Cecilia Poletto. 2012. Contrastive linguistics and micro-variation. The role of dialectology. *Languages in contrast*, 12(1):47–68.

Maria Polinsky. 1995. Cross-linguistic parallels in language loss. *Southwest journal of linguistics*, 14(1-2):87–123.

Ian Roberts. 2019. *Parameter Hierarchies and Universal Grammar*. Oxford University Press, Oxford.

Laurentia Schreiber. 2024. *A (contact-)grammar of Romeyka*. Ph.D. thesis, Otto-Friedrich-Universität, Bamberg.

Giuseppina Silvestri. 2013. *The nature of genitive case*. Ph.D. thesis, Università di Pisa., Pisa.

Giuseppina Silvestri. 2016. Possessivi e partitivi nei dialetti italo-romanzi dell'Area Lausberg. *La lingua italiana. Storia, struttura, testi*, XII:127–142.

Giuseppina Silvestri. 2018. Word-final schwa licensed by prosody and syntax. Evidence from Southern Italian dialects. *Rivista Di Grammatica Generativa*, 3:1–27.

Giuseppina Silvestri. 2020. Possessives in indefinite nominal phrases. A comparison between Italo-Romance and Daco-Romance. *Moderna språk*, 3(114):161–197.

Ioanna Sitaridou. 2014. The Romeyka infinitive. Continuity, contact and change in the Hellenic varieties of Pontus. *Diachronica*, 31(1):23–73.

Peter H. A. Sneath and Robert R. Sokal. 1973. *Numerical taxonomy. The principles and practice of numerical classification*. W. H. Freeman, San Francisco.

Robert Sokal and James Rohlf. 1995. *Biometry*. Macmillan.

Robert R. Sokal and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38:1409–1438.

John Bassett Trumper. 1997. Calabria and Southern Basilicata. In Martin Maiden and Mair Parry, editors, *The dialects of Italy*, pages 355–364. Routledge.

John Bassett Trumper and Marta Maddalon. 1988. Converging divergence and diverging convergence. The Dialect-Language Conflict and Contrasting Evolutionary Trends in Modern Italy. In Peter Auer and Aldo di Luzio, editors, *Variation and Convergence. Studies in Social Dialectology*, pages 217–259. De Gruyter.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 'T Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene Van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan Van Der Lei, Erik Van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Can LLM Agents Identify Spoken Dialects like a Linguist?

Tobias Bystrich[♡] **Lukas Hamm**[♡] **Maria Hassan**[♡]
Lea Fischbach[♢] **Lucie Flek**[♣♠] **Akbar Karimi**[♣♠]

[♡]Department of Computer Science, University of Bonn, Germany
[♢]Research Center Deutscher Sprachatlas, Marburg University, Germany
[♣]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany
[♠]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
s5tobyst@uni-bonn.de ak@bit.uni-bonn.de

Abstract

Due to the scarcity of labeled dialectal speech, audio dialect classification is a challenging task for most languages, including Swiss German. In this work, we explore the ability of large language models (LLMs) as agents in understanding the dialects and whether they can show comparable performance to models such as HuBERT in dialect classification. In addition, we provide an LLM baseline and a human linguist one. Our approach uses phonetic transcriptions produced by ASR systems and combines them with linguistic resources such as dialect feature maps, vowel history, and rules. Our findings indicate that, when linguistic information is provided, the LLM predictions improve. The human baseline shows that automatically generated transcriptions can be beneficial for such classifications, but also presents opportunities for improvement. Code is available on GitHub[1].

Keywords: Spoken dialect classification, German dialects, LLM agents

1. Introduction

The capabilities of Large Language Models (LLMs) have advanced substantially across a wide range of linguistic tasks (Şahin et al., 2020; Srivastava et al., 2023; Chi et al., 2025). However, their performance in low-resource and linguistically grounded tasks, such as spoken dialect identification, remains underexplored. Understanding dialectal variation is important both for linguistic theory, which seeks to explain how language varieties diverge, and for speech technology, where dialect awareness can improve recognition and translation systems. A professional linguist can often identify an unfamiliar dialect by reasoning over linguistic cues or by consulting phonetic transcriptions, which reveal systematic variation through isoglosses on dialect maps (Wolk and Szmrecsanyi, 2018; Tavakoli et al., 2019; Lameli et al., 2020; Lameli, 2022). While some have utilized acoustic modeling (Fischbach et al., 2025a) or text-based methods (Dolev et al., 2024; Peng et al., 2024) for dialect analysis, whether modern LLMs can differentiate between dialects using phonetic representations is under-explored.

In this paper, we investigate whether LLMs can identify Swiss German dialects when provided with automatic phonetic transcriptions of dialectal speech as textual input. Specifically, we ask: *Can LLMs and LLM agents use phonetic patterns to infer dialect identity, and how does this capability compare to that of human linguists?* We use a corpus of automatically transcribed dialectal recordings (Dogan-Schönberger et al., 2021; Plüss et al., 2023) and evaluate the models on dialect classification. We find that, for binary classification of dialects, fine-tuning an encoder-based model still outperforms the LLM approaches. However, the agentic approach considerably improves upon a single LLM model.

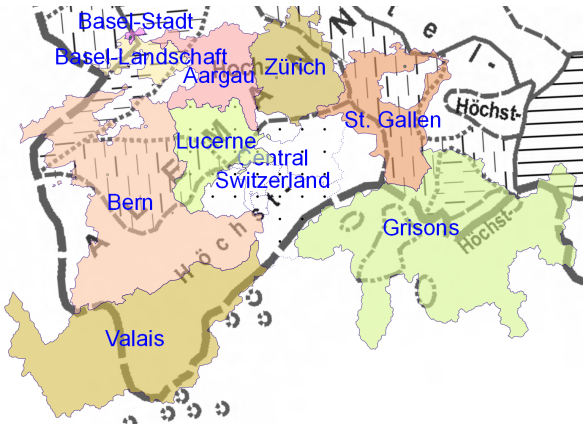


Figure 1: Graphic for SwissDial classes and Innerschweiz made with REDE-SprachGIS and dialect region background (Besch et al., 2008).

Our contributions are: (1) a manual examination of LLM dialect identification using phonetic transcriptions, uncovering how models justify dialect distinctions; (2) an encoder-based model baseline for dialect classification to compare performance with a large language model; (3) comparing the linguistic capabilities of GPT-4o mini and GPT-5, high-

[1] https://github.com/caisa-lab/dialect-classification-with-llm-agents

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 83–92
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

lighting their respective strengths and limitations. These contributions shed light on the limits of LLMs in dialect identification and how far current models approximate human experts when grounded in phonetic representations.

2. Related Work

Large language models (LLMs) have recently demonstrated strong capabilities in linguistic tasks. Foundation models such as GPT-4 (OpenAI, 2023) and multi-agent frameworks like LangGraph (Duan and Wang, 2024) have been used to implement reasoning pipelines, incorporating techniques such as chain-of-thought prompting (Wei et al., 2022) and tool-using agents (Schick et al., 2023). These approaches have shown that LLMs can engage in stepwise inference and apply symbolic reasoning to linguistic tasks, including syntactic analysis, translation, and semantic inference. Despite these advances, the capabilities of LLMs have not been systematically examined in the context of spoken dialect identification. Works such as Wolk and Szmrecsanyi (2018), Tavakoli et al. (2019), Lameli et al. (2020), and Lameli (2022) demonstrate how phonetic transcriptions and dialect maps reveal systematic regional variation through isoglosses, enabling expert linguists to infer dialect affiliation from phonological cues. Fischbach et al. (2025a) investigate acoustic modeling while Dolev et al. (2024) and Peng et al. (2024) utilize text-based methods for dialect classification. The International Phonetic Alphabet (IPA) (Association, 1999) encodes fine-grained pronunciation details and has been widely used in phonetic automatic speech recognition (ASR) models. A number of studies have explored phonetic ASR systems (Xu et al., 2022; Li et al., 2020). Taguchi et al. (2023) introduce MultIPA, a language-agnostic transcription model trained on carefully selected languages to mitigate irregular correspondences between the orthography and the phones. Evaluation of phonetic ASR typically relies on phone error rate (PER), complemented by feature-based and linguistic evaluations (Bystrich, 2025).

Encoder-based speech models form a complementary line of work. For instance, HuBERT (Hsu et al., 2021) performs strongly on dialect classification (Sullivan et al., 2023) and related ASR and phone(me) recognition tasks (Yang et al., 2021), but these models capture primarily acoustic similarity rather than linguistic knowledge. As a result, we explore the capabilities of LLMs in spoken dialect identification from phonetic representations, offering a comparison between encoder models, generative models and a human linguist.

3. Dataset Preparation

While the LLM agent evaluation requires an evaluation set, the HuBERT model needs both fine-tuning and evaluation sets. Therefore, after introducing our utilized datasets in this section, we describe the preparation process.

3.1. Corpora

We use two Swiss German datasets: **SwissDial** (Dogan-Schönberger et al., 2021) and **STT4SG-350** (Plüss et al., 2023) referred to hereafter as **STT**. **SwissDial** contains recordings and manual transcriptions in vernacular orthography for eight dialects: Aargau (AG), Bern (BE), Basel (BS), Grisons/Graubünden (GR), Lucerne/Luzern (LU), St. Gallen (SG), Valais/Wallis (VS), and Zürich (ZH). Each dialect is represented by one speaker who translated sentences from Standard German into the local dialect. The material covers various topics, including news, Wikipedia articles, weather reports, and short stories. **STT** is a substantially larger dataset comprising approximately 343 hours of recordings from 316 speakers across seven dialect regions. Speakers were instructed to read Standard German sentences and render them in their own dialect. The dataset includes metadata on speaker age, gender, and origin and was used under license for this study. While STT provides a broad and demographically diverse sample, making it more representative overall, SwissDial is smaller, with only one speaker per dialect, but it offers richer orthographic details. Together, the two datasets cover a wide range of dialectal and topical variation, including domains such as politics, news, science, and everyday life.

3.2. Label Creation for STT Dataset

To make the classes more consistent across both datasets, we mapped the labels from STT to the 8 SwissDial labels. While Aargau, Lucerne and St. Gallen are not among the dialect region labels in STT, we approximated these by using the existing dialect regions in conjunction with the canton (see Figure 1). We approximated Aargau by using the dialect regions of Zürich and Bern in conjunction with the canton Aargau. We expect that this class cannot easily be predicted since multiple dialect regions occur in the canton Aargau. For the class Luzern/Lucerne, a simpler and likely accurate approximation was possible. The Lucerne dialects are grouped among the STT label ”Innerschweiz” and can further be limited to the canton Lucerne. For St. Gallen, we took the segments labeled with ”Ostschweiz” and the canton St. Gallen.

# Train	Accuracy	# Class predictions		Macro-F1	Accuracy per class	
		High	Highest		High	Highest
400	66.25%	13	67	61.91%	32.5%	100%
4000	**66.25%**	39	41	**66.25%**	65%	67.5%

Table 1: HuBERT baseline for the binary task with best hyperparameters on test data from SwissDial

# Train	Accuracy	Most / least class predictions		Macro-F1
800	11.25%	SG (35)	BE (0)	8.67%
8000	**26.25%**	BS & GR (20)	LU (2)	**22.85%**

Table 2: HuBERT baseline for the 8-class task with best hyperparameters on test data from SwissDial. Abbreviations for dialects: St. Gallen (SG), Bern (BE), Basel (BS), Grisons/Graubünden (GR), Lucerne/Luzern (LU).

Setting	AG	BE	BS	GR	LU	SG	VS	ZH
8-class (800)	40%	0%	10%	0%	0%	30%	10%	0%
8-class (8000)	0%	0%	70%	50%	10%	40%	20%	20%

Table 3: Accuracy per class for HuBERT (8-class) with best hyperparameter configuration on test data from SwissDial. Abbreviations for dialects: Aargau (AG), Bern (BE), Basel (BS), Grisons/Graubünden (GR), Lucerne/Luzern (LU), St. Gallen (SG), Valais/Wallis (VS), and Zürich (ZH).

3.3. Dataset Splits

To prepare the training and test data, all audio files were converted to MP3 format with a sampling rate of 16 kHz. The data were then sampled and divided into three splits each for training, validation, and testing, ensuring that each split contained an equal number of instances for all labels. Following preliminary experiments, all subsequent training sets (for fine-tuning HuBERT) were drawn exclusively from STT, while test and validation sets were created for both STT and SwissDial. Since STT already provides three independent splits, we sampled from these to construct our training, validation, and test sets. For SwissDial, which is a parallel corpus, we ensured that no sentence overlap occurred across the different splits. Because hyperparameter searches and overfitting control are computationally demanding, we prepared both small and large training sets for the 8-class and 2-class classification tasks. For the 8-class task, training set sizes were 800 (small) and 8000 (large); for the 2-class task, 400 and 4000 samples were used, respectively. The small sets were employed for hyperparameter tuning, workflow development, and stability testing, while the large sets were each used for a single full training run. Test splits were fixed to 80 segments due to the high inference time required by the agent. Validation sets followed the same procedure. Although the STT4SG-350 dataset includes speakers from Aargau, they are distributed across multiple broader dialect regions (Basel, Berne, Central Switzerland, and Zurich) rather than being categorized as a standalone set. To specifically evaluate this dialect, the validation set was limited to 80 segments, and 10 additional segments identified as Aargau-origin were sampled from the original training data to ensure a representative sample.

3.4. Classification Setup

Since the eight classes in the SwissDial dataset (8-class problem) were found to be too challenging and not all linguistically defined, we define a 2-class problem. Results for the binary and 8-class problems are shown using the HuBERT baseline in Tables 1 and 2 as motivation for designing a 2-class problem. As we can see in Table 2, the overall performance for the 8-class problem is way below the binary one. Considering the individual classes in Table 3, we can see that many classes are not recognized by this model.

3.5. Conversion into 2-class Problem

Faced with lower accuracies and higher overfitting for this problem than expected, we developed a simpler classification problem. Previous work such as Fischbach et al. (2025b) excluded dialect transition areas and predicted the traditional German dialect regions using a Wiesinger-based map (Besch et al., 2008) as a reference. We take this as a justification for using the dialect regions from Besch et al. (2008), High Alemannic and Highest Alemannic,

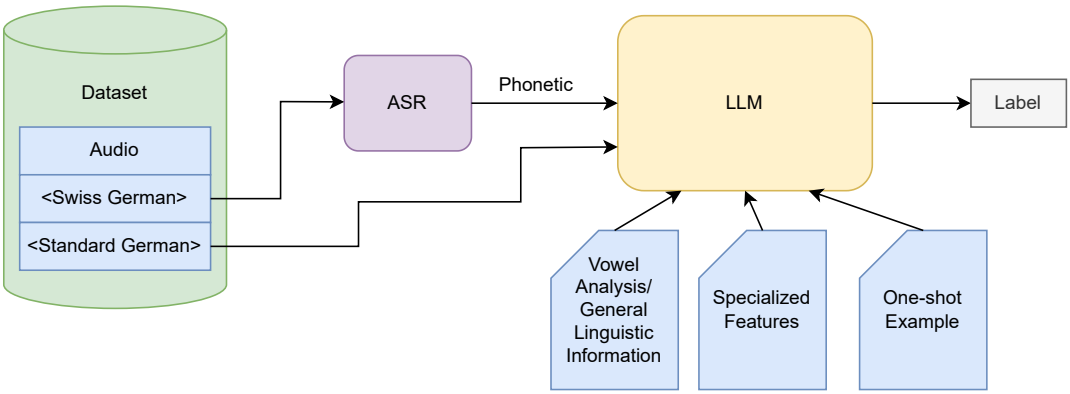


Figure 2: The agentic framework for dialect analysis

and removing transition areas such as Low Alemannic and parts of High and Highest Alemannic.

We used the REDE project SprachGIS[1] to determine the exclusions and mappings to the 2-class problem. Therefore, the new class High Alemannic can only contain Zürich, Aargau and Lucerne. Aargau contains Bernese dialect, however, we excluded Bern as a political unit due to its large overlap with Highest Alemannic (see Figure 1). The canton Lucerne, having some overlap with a transition area and minor overlap with Highest Alemannic, is the least neatly mapped inclusion. We decided in favor of its inclusion since it seems to have considerably smaller overlap than the excluded classes and since not including it might limit its ability to actually classify High Alemannic given a high bias towards Zürich.

For Highest Alemannic, the overlap of Grisons/-Graubünden with the High Alemannic region made its exclusion important (see Figure 1). This only left Valais for SwissDial. For STT, increasing the variance was possible by using Innerschweiz, limited to cantons totally situated in the Highest Alemannic region. We validated this choice by comparing the performance on an STT vs. a SwissDial test set (see Section 5).

All 2-class datasets have an equal number of segments for High and Highest Alemannic, while the source classes (Aargau, Lucerne, Zürich; Valais, Innerschweiz_Highest) get an equal number within their class, barring a difference of 1 due to discrete numbers. Innerschweiz_Highest is not used for the 8-class problem.

4. Methodology

Figure 2 shows the design for our dialect agent. In this section, we describe its components.

4.1. ASR Model

We used the XLSR-53 version[2] of Wav2Vec2Phoneme (Xu et al., 2022) for phonetic transcription. The reported phone error rate (PER) for zero-shot transcription without an LLM is 33.3% (Xu et al., 2022). To answer how well a human linguist can classify dialects based on these transcriptions, we perform an evaluation with human baseline in Section 6.

4.2. Base Prompt and Query

Base Prompt Box

You are now a linguist who needs to identify dialects based on feature descriptions and observations. As a linguist, linguistic reasoning is far more important than coding, coding is likely unnecessary. You are provided transcriptions in the IPA that were generated automatically as well as translations into Standard German to help you interpret the dialect transcription. Please consider the topics with a focus of a linguist, a dialectologist and a phoneticist. Vowels and consonants must always be considered in their word. It makes no sense to just check for the presence of specific phones. Instead, they must be considered in relation to the morpheme in question. **[THIS PART IS FOR BINARY PROBLEM]** Your first task is to identify a Swiss German dialect into one of two dialect regions: High Alemannic, Highest Alemannic. Please output as your final reply only the name of the dialect region. **[THIS PART IS FOR 8-CLASS PROBLEM]** Your first task is to identify a dialect as one of several Swiss German dialects. Please output your reply as the two letter short form (ag, be, bs, gr, lu, sg, vs, zh) for (Aargau, Bern, Basel, Grisons/Graubünden, Lucerne/Luzern, St. Gallen, Valais/Wallis, Zürich).

[1] https://regionalsprache.de

[2] https://huggingface.co/docs/transformers/model_doc/wav2vec2_phoneme

[v eː r h aɪ t d iː v oː n n ɑː b ɔ t l ə r
a n iː r ə ʃ l ɛ t ʃ t ə a r b aɪ t s t ɔ k
p s y ɛ x t ʊ n ə f iː r ɪ ʃ p ʊ l iː t
ɪ ʃ l aʊ f b a n ts r uː k l ɛ k t]

Figure 3: Sample transcription with Highest Alemannic features. Orthographic transcription: *Wir haben Yvonne Beutler an ihrem letzten Arbeitstag besucht und auf ihre politische Laufbahn zurückgeschaut* [*-gelugt*]

The base prompts for both problems can be found in Base Prompt Box. The boldfaced sentences are not part of the prompt, but only to show which part was used for which problem. The query we provide for the model for classification is preceded by "[USER]" and consists of the phonetic transcription in the IPA and a Standard German translation. Figure 3 shows a sample transcription.

4.3. Linguistic Information

As input, we provide the model with linguistic information to enable it to go beyond its pretrained knowledge. Some of this information include the dialect features derived from maps since the model might not be able to recognize distinguishing features of Swiss German dialect regions from IPA transcriptions. In addition, we feed the model an explanation of the linguistic information in plain English. To further improve the accuracy, we also provide historical vowel information, since vowel sound changes are highly important. This is provided as a table. We give the model IPA charts for vowels and consonants, to help achieve a more definition-consistent IPA interpretation of the model. Finally, we provide the model with a sample reference evaluation so that it can learn from the thought process of a linguist.

4.4. LLM

We used OpenAI's GPT-4o mini via API. We set the temperature to zero to maximize truthful outputs and minimize unpredictable factors. Since non-mini GPT versions by OpenAI such as GPT-5 could not be used via API, we randomly compared some outputs via ChatGPT with GPT-5 as the backbone.

4.5. LangGraph

We used a LangGraph agent for our dialect classification task. It uses the same linguistic information and prompt but breaks it down into different nodes. Our system consists of 2-3 nodes where each of them uses a different prompt to send requests to the LLM (GPT-4o mini). We structured the LangGraph agent using multiple **nodes**. The node **vowel and consonant analysis** focuses on vowel- and consonant-related cues and returns an object with per-class confidence scores and brief reasoning. And the **specialized features analysis** considers broader phonological features and returns an object with class probabilities and a final prediction. The state object passed between nodes includes inputs (audio filename, ASR transcription, Standard German transcription), intermediate analysis (vowel analysis and dialect features analysis results).

5. Experiments and Results

We tested all of the models with balanced 80 test samples from SwissDial. The HuBERT Baseline was trained and validated on STT to avoid overfitting. We used a balanced small training split of 400 segments for the hyperparameter search and then trained with a balanced training split of 4000 segments using the same hyperparameters. As an additional validation for HuBERT, we additionally tested it on a STT split to compare the difference to the SwissDial data.

5.1. Baselines

LLM Baseline Our LLM baseline is similar to Figure 2 without the additional linguistic information, with only the data and the phonetics from the ASR model as inputs. It has the same model and base prompt as the main dialect agent, intended as a fair comparison with the encoder-based classifier and to show whether our additional information can improve the LLM's performance.

HuBERT HuBERT uses a training approach similar to that of BERT (Devlin et al., 2019). The raw audio signal is first segmented into short frames of approximately 20–25 ms. These frames are then represented by Mel-frequency Cepstral Coefficients (MFCC) features, which are clustered via k-means into up to 500 groups. The resulting cluster assignments serve as pseudo-labels (cluster IDs) for self-supervised pre-training.

During training, the audio frames are processed by a convolutional waveform encoder, producing latent representations. After parts of these representations are masked, they are fed into a Transformer network (Vaswani et al., 2017). The Transformer is trained to predict the original cluster IDs of the masked segments, thereby learning robust and context-aware speech representations.

For classification, we used a two-layer head, consisting of 128 units in the first GELU layer and

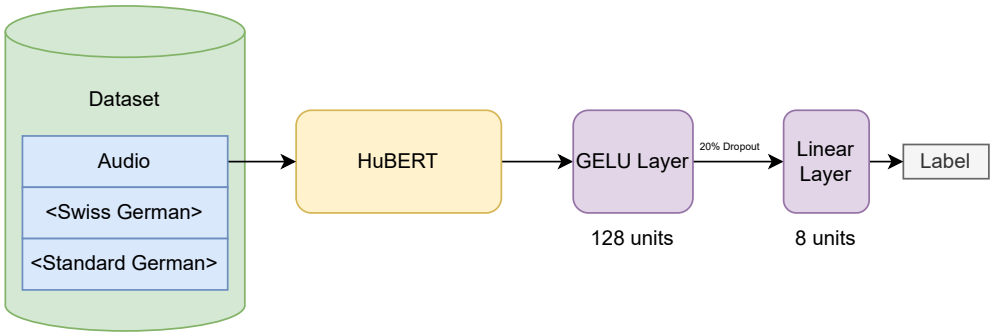


Figure 4: HuBERT baseline architecture

lr	Batch	Accuracy	# Class predictions		Macro-F1	Accuracy per class	
			# High	# Highest		High	Highest
10^{-3}	8	52.5%	61	19	50.99%	70%	35%
	16	46.25%	61	19	42.27%	72.5%	20%
	32	56.25%	25	55	54.66%	37.5%	75%
	64	62.5%	22	58	60.5%	40%	85%
10^{-4}	8	50%	76	4	37.3%	95%	5%
	16	53.75%	67	13	47.8%	87.5%	20%
	32	47.5%	70	10	38.9%	85%	10%
	64	**66.25%**	**13**	**67**	**61.91%**	**32.5%**	**100%**

Table 4: Results of hyperparameter search for the HuBERT baseline on test data from SwissDial.

Experiment Configuration	**Accuracy (%)**
Base Prompt Only (2 runs average)	47.8
Specialized Features Analysis Node	47.5
Vowel & Consonant Analysis Node	55.0
Vowel & Consonant Node + Specialized Features Node (2 runs average, no IPA explanations from GPT)	**58.0**

Table 5: Accuracy of different LangGraph experiment configurations. We use the best results for comparison with other methods.

8 units in the final linear output layer. Additionally, a dropout rate of 0.2 was applied between the first and last classifier-head layer. The number of training epochs was determined using early stopping. We set a minimum threshold of 3000 training examples, translated into the corresponding number of epochs, to allow for initial mistakes in the early stages. Beyond this threshold, training was stopped once the development loss exceeded the lowest dev loss observed across all epochs. Figure 4 shows the diagram for the HuBERT baseline.

We used the HuBERT model to compare the agentic approach to more conventional speech classification approaches. Using the multilingual version mHuBERT-147 (Boito et al., 2024) ensured that Germanic languages were included during pretraining. After doing a hyperparameter search shown in Table 4, we choose the best results for comparison with other methods (lr = $1e{-}4$, batch size = 64).

Human Linguist The last binary test set of 80 segments was analyzed by one of the authors with linguistic experience based on the same linguistic information that the LLM is provided with and the same auto-generated transcriptions. They used only the same linguistic information, aiming to avoid any personal opinion. If the evidence did not suffice for determining a class, the prediction was counted as neither correct nor incorrect to avoid any biases.

5.2. LangGraph Findings

We investigated various LangGraph configurations to study the impact of prompt engineering and amount of information provided. Experiments revealed that using only the base prompts gives an average accuracy of 47.8 percent (Table 5). However, when the additional vowel & consonant and specialized features nodes are added to the setup, accuracy improves to 58% on average and 62.5% for the best run. A prediction imbalance was also observed which was that for an 80 sample dataset with an even 40/40 split for a 2-class problem, the

Model	Accuracy	# Class predictions		Macro-F1	Accuracy per class	
		High	Highest		High	Highest
HuBERT	66.3%	39	41	66.3%	65%	67.5%
LLM	47.8%	30	50	47%	35.6%	60%
Human	72.5%	33	46	72.3%	80%	65%
Agentic	58%	23	57	56.3%	37.5%	78.8%

Table 6: Results for the 2-class problem

Manual Transription	**ChatGPT (GPT-5)**	**GPT 4o-mini**
[ ] Die [overlaps with next word]	[d oː] Die	[d i] Die
[d oː t r ʊ] da[d]raus [overlaps with next word]	[t r ʊ] daraus	[d aː r ʊ] daraus
[ʃ l ɪ p ə n d] schlüpfenden	[ʃ l ɪ p ə n d] schlüpfenden	[ʃ l ʊ p f ə n d] schlüpfenden
[r aʊ p ɑː] Raupen	[r aʊ p ɑː] Raupen	[r aʊ p ɑː] Raupen
[ɛ r m ɛ r ə n] ernähren	[ɛ r m ɛ r ə n] ernähren	[ɛ r n ɛː r ə n] ernähren
[z ɪ x] sich	[z ɪ x] sich	[z ɪ x] sich
[f oː d ɛ r ɜ] von der	[f oː d ɛ r] von der	[f oː n] von [d iː s ə] dieser
[p f l a n ts ə] Pf anze	[ɜ p f l a n ts ə] Pf anze	[p f l a n ts ə] Pf anze

Figure 5: Alignments between human baseline and GPT models

model predicted 57 instances of ”Highest” and only 23 instances of ”High”.

5.3. Overall Comparison

Table 6 shows the results for the baseline models, agentic framework, and the human baseline for the 2-class problem. The HuBERT baseline model gives an accuracy and macro-F1 of 66.3% for the 2-class problem. Compared to the other classifiers, HuBERT achieved the best macro-F1 and the second-best accuracy after the human baseline, which makes it the highest-achieving non-human model in this metric. The most and least predicted classes show well-distributed predictions (41/39) that are close to the original 50/50 distribution of the ground truth test set. For the LLM-based approaches, two runs were averaged with the standard deviation of 4.1 percentage points for the LLM baseline and 4.4 for the LangGraph agent. The agentic approach shows a higher performance than the LLM baseline, achieving an improvement of 10.2 percentage points. The human baseline achieved the highest overall accuracies.

6. Discussion on Linguistic Approaches

This section discusses the results for the agent, the human baseline and the LLM baseline. The LLM baseline shows results around 50%, suggesting it does not have adequate previous knowledge for this task. The reported performance is the average of two test runs. For the human linguist baseline, the analysis of the 80 SwissDial segments provided enough evidence for a decision in 58 cases. For these cases, the human accuracy was 81%. There were 12 ambiguous segments for Highest Alemannic and 10 for High Alemannic. As a result, for the calculation of these predictions as neutral (to avoid any biases), we assigned half of each with the correct and half of each with the incorrect value. This led to an overall accuracy of 72.5%. One could also allow the LLM to have a third option to abstain from choosing either class. However, early tests showed that this option may lead to a worse performance.

This result serves as our extrinsic evaluation for the ASR model, showing that even with the sub-optimal phone error rate, feeding only one sentence at a time can still allow for a classification of High and Highest Alemannic. Improving the ASR quality would likely further enhance the human accuracy.

6.1. Comparative Analysis with GPT-5

ChatGPT showed more robustness concerning the work with phonetic transcription, as we will show using an example. The alignment of phones to Standard German is an important step for the linguistic analysis since the analysis of the sound changes and morphological information depends on the phones being aligned to the correct etymons. It can be summarized that GPT-4o mini fails in ways that are difficult to conceive for a human, even in

normal cases, while ChatGPT showed surprising skill even in challenging situations (see Figure 5). Here we can see that GPT-5 aligned the words almost perfectly. It was able to handle this challenging case with multiple overlapping words quite well. However, GPT-4o mini hallucinated the transcriptions.

6.2. Other Challenges

The dialect recognition with linguistic resources was likely affected by ASR performance, however, the results of the extrinsic evaluation by a human surpass the HuBERT baseline. Additionally, Standard German loan words into Swiss German made the classification more difficult since they don't fit the dialect's sound changes. Lastly, some sentences may not contain any words with the potential to identify a dialect without ambiguity.

7. Conclusion

We implemented an LLM agent to see if it can utilize linguist information about dialects as context to classify dialects (of Swiss German) given phonetic transcriptions and translation into the literary language (Standard German). We focused on dialect regions, instead of towns' dialects, and excluded dialect transition areas. Implementing a LangGraph and GPT-4o-mini-based agent showed a considerable improvement over a baseline LLM. As a result, we conclude that LLM agents can, to some extent, use linguistic information and phonetic transcriptions to improve their dialect understanding. However, given the difficulty of the task as well as the quality of the ASR models, they would need more supervision to show human-level capabilities in dialect analysis.

8. Limitations

Given our limited budget, we used GPT-4o mini in our experiments. While it shows considerable capacity for our task, one could also use more capable models such as GPT-4 or GPT-5 as well as other model families such as Gemini (Team et al., 2023) and Llama (Touvron et al., 2023). In addition, ineffective tool use of the agent necessitated more focus on prompt engineering, something that can be investigated more for an improved framework. Finally, with the availability of more domain experts, one could increase the number of annotated test samples to perform a more comprehensive analysis of our system.

9. Ethical Considerations

We ensured that the data is used responsibly. Speaker identities were anonymized in the released datasets. Potential misuse of speaker identification technology could threaten privacy; we recommend deployment only in contexts with explicit user consent and appropriate legal frameworks.

Acknowledgments

This research is supported by the Academy of Science and Literature Mainz (grant REDE 0404), the German Federal Ministry of Education and Research (BMFTR) (grant AnDy 16DKWN007) and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B and the Research Center Deutscher Sprachatlas Marburg.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Werner Besch, Ulrich Knoop, Wolfgang Putschke, and Herbert E. Wiegand. 2008. *Dialektologie. 2. Halbband*. De Gruyter.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model.

Tobias Bystrich. 2025. Multilingual Automatic Phonetic Transcription – a Linguistic Investigation of its Performance on German and Approaches to Improving the State of the Art. Master's thesis, University of Bonn.

Nathan Andrew Chi, Teodor Malchev, Riley Kong, Ryan Andrew Chi, Lucas Huang, Ethan A Chi, R Thomas McCoy, and Dragomir Radev. 2025. ModeLing: A Novel Dataset for Testing Linguistic Reasoning in Language Models. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 105–114.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Pelin Dogan-Schönberger and Julian Mäder and Thomas Hofmann. 2021. *SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German*.

Eyal Dolev, Clemens Lutz, and Noëmi Aepli. 2024. Does Whisper Understand Swiss German? An Automatic, Qualitative, and Human Evaluation. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 28–40.

Zhihua Duan and Jialin Wang. 2024. Exploration of LLM Multi-Agent Application Implementation Based on LangGraph+CrewAI.

Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. 2025a. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Proc. Interspeech 2025*, pages 2780–2784.

Lea Fischbach, Caroline Kleen, Lucie Flek, and Alfred Lameli. 2025b. Does Preprocessing Matter? An Analysis of Acoustic Feature Importance in Deep Learning for Dialect Classification.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.

Alfred Lameli. 2022. Syllable Structure Spatially Distributed: Patterns of Monosyllables in German Dialects. *Journal of Germanic Linguistics*, 34(3):241–287.

Alfred Lameli, Elvira Glaser, and Philipp Stöckle. 2020. Drawing Areal Information from A Corpus of Noisy Dialect Data. *Journal of Linguistic Geography*, 8(1):31–48.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, and Florian Metze Alan W Black. 2020. Universal Phone Recognition with a Multilingual Allophone System.

OpenAI. 2023. GPT-4 Technical Report.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! Recognizing Named Entities in Bavarian Dialectal Data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493.

Michel Plüss, Jan Milan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardzic, Manfred Vogel, et al. 2023. Stt4sg-350: A speech corpus for all swiss german dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the Robustness of Arabic Speech Dialect Identification.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal Automatic Phonetic Transcription into the International Phonetic Alphabet.

Shahin Tavakoli, Davide Pigoli, John AD Aston, and John S Coleman. 2019. A Spatial Modeling Approach for Linguistic Object Data: Analyzing Dialect Sound Variations Across Great Britain. *Journal of the American Statistical Association*, 114(527):1081–1096.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35:24824–24837.

Christoph Wolk and Benedikt Szmrecsanyi. 2018. Probabilistic Corpus-based Dialectometry. *Journal of Linguistic Geography*, 6(1):56–75.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I. Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing Universal PERformance Benchmark.

Beyond Accuracy: Analyzing Dialect Confusion in Automatic Speech-Based Dialect Classification

Lea Fischbach[1], **Alfred Lameli**[1], **Lucie Flek**[2,3]
[1]Research Center Deutscher Sprachatlas, Marburg University, Germany
[2]Lamarr Institute for ML and AI, Germany
[3]b-it Center, University of Bonn, Germany
lea.fischbach@uni-marburg.de

Abstract

Automatic dialect classification is commonly treated as a supervised task with a primary focus on overall accuracy. In this paper, we argue that classification errors and model uncertainty provide valuable insights into dialectal structure and variation. We analyze a speech-based dialect classification model trained on German dialect data from three generations and evaluated across 250 speaker-disjoint splits (median weighted $F_1 = 0.42$). A systematic confusion analysis shows that misclassifications are largely explained by speaker diversity, dialectal similarity, geographical proximity, and speaker self-assessment. Among these factors, the number of speakers per dialect has the strongest impact on performance, while frequent confusions between closely related dialects reflect inherent linguistic similarity rather than model limitations. Generational analyses further indicate that younger speakers exhibit reduced dialectal distinctiveness, although core dialectal features remain shared across generations. By explicitly modeling classification uncertainty, the proposed approach enables the analysis of dialect transition areas and gradient dialect boundaries. Overall, this work demonstrates that automatic dialect classification can serve not only as a predictive task but also as a tool for dialectological analysis.

Keywords: automatic dialect classification, dialect confusion, dialect continua, generational variation

1. Introduction

Automatic dialect classification from speech is commonly framed as a supervised classification task (Hosseini-Kivanani et al., 2025; Ghafoor et al., 2025). While this perspective emphasizes overall performance, a closer examination of classification errors, confusions, and low-confidence predictions reveals important information about dialectal structure, speaker variation, and the gradient nature of dialect boundaries.

In line with this view, several studies have pointed out structural challenges inherent to dialect classification. Jokisch and Dobbriner (2019), for example, classified nine German dialects using acoustic features such as MFCCs. They report that some dialects and dialect groups are inherently similar and therefore difficult to distinguish, even for human listeners. In addition, they note that some speakers in their test corpus articulate close to Standard German, which may conflict with the forcibly assigned dialect labels used during training. Further challenges include unequal speaker distributions across dialects, unbalanced dialect region sizes, and sparse training data for certain dialect groups. While these factors are clearly identified, their individual impact on misclassification behavior is not examined in detail.

Similarly, Stucki and Randjelovic (2021) investigated the automatic classification of 21 Swiss German dialects using Wav2Vec-XLSR-53 representations. They observed that some dialects were recognized reliably, whereas others were almost entirely misclassified. Moreover, misclassified samples were predominantly assigned to geographically neighboring dialects, indicating dialectal similarity. At the same time, they report that a small number of speakers contributed a large proportion of samples for some dialects, raising the risk that the model learned speaker-specific characteristics rather than dialect features.

In addition, Misganaw and Roller (2022) classified 7,814 short audio snippets from eight dialects in Middle West Germany using a multilabel SVM. However, their dataset comprised only one to two speakers per dialect, with the same speakers used for training and testing. Under these conditions, high performance likely reflects speaker recognition rather than dialect classification, a factor that was not explicitly addressed in their study.

Taken together, these studies demonstrate that reported classification results are strongly influenced by the choice and similarity of dialects and dialect groups, the number and distribution of speakers, and the degree to which speaker-specific variation is controlled (Darjaa et al., 2018). Despite these observations, systematic analyses that disentangle the relative contributions of these factors to misclassification behavior remain rare, particularly with respect to quantitatively separating speaker-related effects from dialectal similarity under speaker-disjoint evaluation settings. While un-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 93–103
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

supervised and data-driven approaches aim to derive similarity structures from acoustic data or predefined linguistic features, for example through distributional similarity metrics such as token distribution similarity (TDS) (Gogoulou et al., 2024) or dialectometric clustering based on linguistic distance measures (Sciarretta, 2024), they do not explicitly model misclassification behavior with respect to predefined dialect categories or allow for a direct analysis of how similarity interacts with speaker variation and data-related factors. In this paper, rather than treating misclassifications solely as model failures, we interpret systematic confusions as potentially informative signals reflecting underlying dialectal relationships, speaker distributions, and sociolinguistic variation. This perspective aligns with previous work on dialect classification, which emphasizes that low performance may arise from dialectal similarity, speaker heterogeneity, and imbalanced data distributions rather than from limitations of the classification model alone (Jokisch and Dobbriner, 2019; Ferragne and Pellegrino, 2007; Stucki and Randjelovic, 2021).

With our study, we therefore do not aim to achieve maximal classification accuracy. Instead, we analyze a speech-based dialect classification model with the goal of understanding systematic misclassifications and low-confidence predictions. We show how such patterns can be related to sociolinguistic factors (generation and speaker self-assessment), geolinguistic properties (dialectal similarity and transition areas), and data-related aspects (speaker number). By doing so, we demonstrate how dialect classification models can be used as analytical tools to investigate dialect variation beyond hard class assignments.

The contributions of this paper are as follows: (i) we present a speech-based dialect classification model trained on a large German dialect corpus; (ii) we provide a systematic analysis of confusions and misclassifications, identifying four main contributing factors; (iii) we introduce an approach to exploit model uncertainty for the analysis of dialect transition areas.

2. Data and Annotation

The audio data used in this study originates from the REDE project (Schmidt et al., 2020–), a large-scale project on regional linguistic variation in Germany. We focus on a subset of the corpus in which speakers were asked to translate 40 Standard German sentences into their local dialect.

The dataset comprises recordings from three generational groups, defined by age: older speakers (60+), middle-aged speakers (42–59), and younger speakers (17–26), with exact age available for 516 speakers (≈90%). The distribution of unique speakers as well as the total duration of audio per generation used in this study are summarized in Table 1. The corresponding age distribution of speakers is shown in Figure 1, illustrating both the separation between the predefined age groups and the variability within them.

Generation	#Speakers	Usable audio (s)
Older speakers	198	63520
Middle-aged speakers	237	58750
Younger speakers	139	31700

Table 1: Number of unique speakers and total duration of usable audio (in seconds) per generational group included in this study.

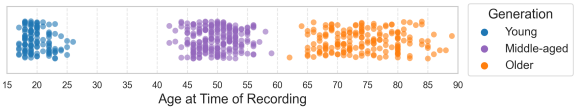


Figure 1: Age distribution of speakers across the three generational groups (older, middle-aged, younger) in the dataset. Each point represents an individual speaker age at the time of recording.

Figure 2 illustrates the geographical distribution of the dialect areas and their transition zones based on Wiesinger (1983), as well as the recording locations covered in the dataset. Dark blue points indicate recordings from core dialects, each associated with a specific dialect label, while light blue points represent recordings from transition zones. These transition zones are of particular interest, as they reflect the gradient nature of dialectal variation rather than strictly discrete boundaries. Dialect areas shown in gray were excluded from the analysis due to an insufficient number of speakers.

2.1. Speaker Self-Assessment

In addition to the speech recordings, speaker self-assessments were collected for each speaker. The self-assessment data were gathered through a standardized questionnaire[1]. Speakers were asked, among others, to rate their active (spoken) competence in the local dialect. The relevant question was formulated as follows: “How well can you speak the [placeholder for the term used by the informants to refer to the way of speaking of long-established residents in the locality] of your hometown?” Responses were provided on a rating scale ranging from 0 (*not at all*) to 6 (*perfect*), with increments of 0.5. The collected data are available in tabular form[2].

[1] available at https://doi.org/10.57712/lingurep-59870
[2] https://doi.org/10.57712/lingurep-59867

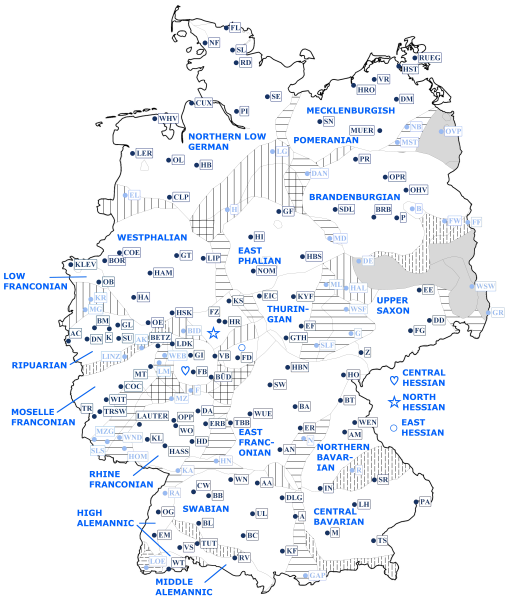


Figure 2: Dialect areas, transition zones according to Wiesinger (1983), and survey locations in the dataset. White areas represent labeled dialect regions, hatched areas transition zones, and gray areas regions excluded from analysis. Points indicate survey locations with location abbreviations.

3. Model Architecture and Training

Figure 3 provides an overview of the important steps from the training pipeline. For model training, only recordings from dialect areas were used. Speech samples originating from transition zones were excluded from the training process and are considered exclusively in the analysis.

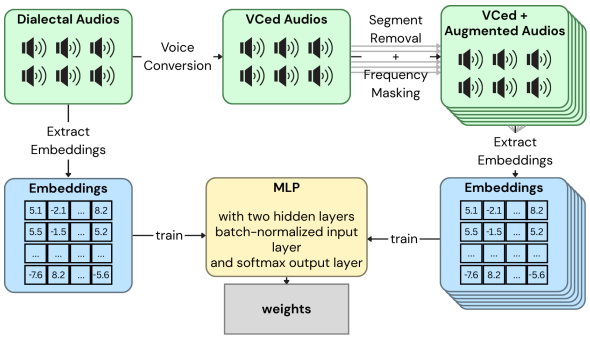


Figure 3: Overview of the training pipeline.

All audio recordings were first subjected to voice conversion, mapping all audios to a single target speaker voice. Voice conversion is not used to improve absolute performance, but to isolate dialect-related variation by minimizing speaker-specific acoustic variability. This step is motivated by the observation that voice conversion reduces speaker-related variability other than dialectal and thereby enables the model to focus more strongly on dialect-specific linguistic and phonetic features (Fischbach et al., 2025a) while preserving linguistic content (Sisman et al., 2020). To further increase data variability and robustness, six augmented versions were generated for each voice-converted audio file using frequency masking and segment removal as described in (Fischbach et al., 2025b).

All resulting audio files were subsequently cut into fixed-length segments of 10 seconds and normalized by resampling to 16 kHz, converting to 16-bit depth, and mixing down to mono. Incomplete final segments shorter than 10 seconds were discarded. Given the total duration per generation reported in Table 1, the number of resulting segments follows directly from this fixed windowing procedure. Speech embeddings were extracted from each segment using Google's TRILLsson 4 model (Shor and Venugopalan, 2022) and used to train a multilayer perceptron (MLP) classifier. The MLP consists of a batch-normalized input layer followed by two fully connected hidden layers with He-normal initialization, L2 kernel and L1 activity regularization, and LeakyReLU activations. Dropout is applied after each hidden layer to reduce overfitting. The output layer uses a softmax activation to predict dialect labels. The model is trained using categorical cross-entropy loss and the Adam optimizer. Early stopping based on validation loss is employed to prevent overfitting, with the best-performing model checkpoint restored.

A speaker-independent data split was employed. For each dialect, 10% of speakers were assigned to the validation set and 10% to the test set, with the remaining speakers used for training. Crucially, each speaker appears in exactly one of the subsets (training, validation, or test), preventing the model from exploiting speaker-specific characteristics and ensuring that performance reflects dialectal generalization rather than speaker recognition.

The complete training pipeline, including preprocessing, augmentation, embedding extraction, and model training, is implemented in a publicly available codebase[3].

4. Overall Classification Performance

Across 250 evaluation runs, the model achieved a mean and median weighted F1-score of 0.42, with a standard deviation of 0.036. These results are based on 15,397 original segments and an additional 46,272 augmented samples. Multiple runs were performed because no single fixed train–validation–test split was considered representative for the present analyses. Instead, repeated speaker-independent splits with randomly assigned

[3] https://github.com/WoLFi22/DialectClassificationPipeline

speakers were used to reduce split-specific bias and to ensure that analyses were not driven by particular test speakers. Importantly, all analyses focus on relative confusion patterns aggregated across evaluation runs. The number of runs was chosen empirically, as the overall results stabilized well before 250 iterations.

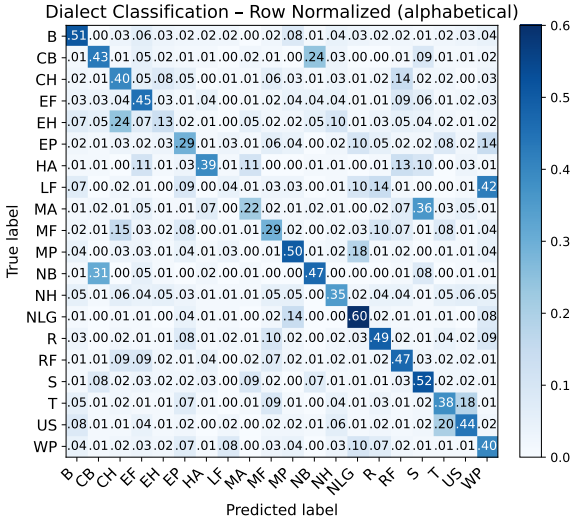


Figure 4: Row-normalized confusion matrix of dialect classification results. Abbreviations: B=Brandenburgian, CB=Central Bavarian, CH=Central Hessian, EF=East Franconian, EH=East Hessian, EP=Eastphalian, HA=High Alemannic, LF=Low Franconian, MA=Middle Alemannic, MF=Moselle Franconian, MP=Meklenburgish Pomeranian, NB=North Bavarian, NH=North Hessian, NLG=Northern Low German, R=Ripuarian, RF=Rhine Franconian, S=Swabian, T=Thuringian, US=Upper Saxon, WP=Westphalian

Figure 4 shows the row-normalized confusion matrix aggregated across all evaluation runs. The matrix indicates systematic asymmetries in misclassification behavior, suggesting that certain dialect pairs are more frequently confused than others. In addition, several dialects show low recall values, indicating that they are rarely predicted correctly.

A detailed analysis of these effects, including generational influences and systematic confusion patterns between dialects, is presented in Section 5 and Section 6.

5. Analysis of Age Groups

Figure 5 shows self-assessed speaking proficiency by birth-year group, based on the questionnaire described in Section 2.1. The birth-year cohorts can be grouped into three generations (older, middle-aged, and younger; cf. gray separators in the figure). Speakers in the older generation (three oldest cohorts) show lower variance and rate their dialect competence about two points higher (on a six-point scale) than the middle-aged generation. At the same time, a clear decrease in self-assessed dialect competence from the middle-aged to the younger generation is visible. Due to the absence of speakers born between 1970–1979, values for this cohort were interpolated solely to provide a continuous visual representation of the generational trend. These patterns in self-assessed competence are reflected in the classification results.

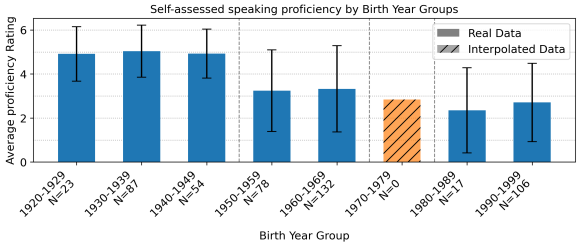


Figure 5: Average self-assessed dialect proficiency by birth-year group. Blue bars show observed data, hatched bars interpolated values, and error bars standard deviation. Gray vertical lines separate the three generations (old, middle-aged, and young).

Classification performance differs markedly across generations. The older group achieves the highest correct classification rate (46.8%), followed by the middle-aged group (45.7%), while the younger group performs markedly worse (28.1%), mirroring the lower self-assessed competence observed above.

The observed trend is consistent with established findings in dialectology. Niebaum and Macha (2014) note that dialect competence is typically correlated with speaker age, arguing that older speakers are more likely to possess dialect competence, use dialect forms more frequently, and produce more distinct dialectal realizations. Similar conclusions are drawn by Lameli (2025), who interprets this pattern as evidence of a diachronic decline in dialect competence and describes it as an intergenerational decrease in dialectal intensity.

To further isolate generational effects, separate classification experiments were conducted for each generation group, using only speakers from that generation for training, validation, and testing. Under this constraint, correct classification rates drop to 21.3% for the younger generation, 40.5% for the middle-aged generation and 39.1% for the older generation. This decrease indicates that all generations benefit from the inclusion of speech data from other age groups. While part of this effect is likely attributable to the increased amount of training data, it is also consistent with the assumption that dialectal features are, to some extent, shared

across generations despite differences in competence and usage frequency.

To further disentangle the role of shared representations from data quantity effects observed in the previous experiment, we conducted a controlled cross-generational comparison. In the baseline setting, training, validation, and testing were all performed on younger speakers. In the cross-generational setting, training and validation were performed on older speakers, while testing was performed on younger speakers. In both settings, the same set of dialects was used. For each dialect, the number of speakers was kept constant across training, validation, and test splits, and matched between the two experimental settings. Under these conditions, classification accuracy drops from 23.5% in the within-generation setting to 15.8% in the cross-generational setting. Note that the within-generation result differs from the previously reported value (21.3%) due to the restriction to a subset of dialects shared across both generations, resulting in a slightly simplified classification task. Despite this substantial decrease, performance remains clearly above chance level, indicating that dialectal features are, to some extent, shared across generations. At the same time, the performance gap suggests systematic differences in the realization of these features.

Taken together, both classification results and self-assessment data consistently indicate a generational decline in dialect competence. While younger speakers still share core dialectal characteristics with older generations, reduced dialect proficiency and weaker realization of dialect-specific features appear to limit classification performance for this group.

6. Dialect Confusion Analysis

While the overall classification performance provides a global assessment of model behavior, it does not explain why certain dialects are systematically misclassified or confused with others. In this section, we therefore analyze confusion patterns at the dialect level in order to identify linguistic, geographic, and data-related factors that may contribute to classification errors.

The four explanatory factors considered in this analysis are the number of speakers, geographical proximity, dialectal similarity, and speaker self-assessment. The number of speakers refers to the number of unique speakers per dialect included in the dataset. Each factor is represented categorically according to the levels defined in Table 2. A check mark (✓) indicates the strongest or most direct connection between two dialects, while a circle (◐) represents intermediate conditions — for example, dialects that are geographically close but separated by a transition zone, or dialects that share only partial similarity as identified in related studies. A cross (✗) denotes the weakest connection or absence of a relationship along the respective factor dimension. The classification thresholds (e.g., speaker count ranges or percentile cutoffs) are explained in the respective analytical sections.

	Factor levels		
Factor	✓	◐	✗
Speakers	<15	15-20	>20
Geo Proximity	adjacent	close	distant
Similarity	proven	partial	none
Self-assess.	<20th perc.	20th perc.-avg	>avg

Table 2: Factor coding scheme used to categorize explanatory variables. The specific thresholds and definitions for each factor are detailed in the corresponding sections of the analysis.

To structure the analysis, we examine two types of cases: (i) dialect pairs that exhibit high confusion rates (>20%), and (ii) dialects with low correct classification performance (<40%). These cases are summarized in Table 3 and form the basis for our investigation of factors that may explain the observed patterns. For confused dialect pairs, the table lists the true and predicted dialects together with the proportion of instances in which the true dialect was predicted as the other. For dialects with low correct classification performance, only the affected dialect is listed, and the reported value corresponds to its correct classification rate. In both cases, the table provides a factor-based coding of potential explanatory variables.

6.1. Number of Speakers

Figure 6 illustrates the relationship between dialect-wise classification performance and the number of speakers per dialect. Classification performance increases with the number of speakers, showing a strong positive correlation between the logarithm of speaker count and the F1-score (Pearson $\rho = 0.86$, $p < 0.001$), which is higher than the corresponding correlation in linear space (Pearson $\rho = 0.79$, $p < 0.001$). A simple linear regression model on log-transformed speaker counts yields an R^2 score of 0.73, indicating that a substantial proportion of performance variance can be explained by speaker diversity alone.

Across dialects, the number of speakers and the total number of segments per dialect are almost perfectly correlated (Spearman $\rho = 0.99$, $p < 0.001$). Nevertheless, previous experiments using voice conversion indicate that classification performance is more strongly influenced by speaker diversity than by the sheer number of segments. This moti-

	True	Predicted	Rate	Reason			
				Spk	Geo	Sim	Self
Confused	Middle Alemannic	Swabian	0.36	✓	✓	◐ Lameli (2013)	✗
	Central Bavarian Northern Bavarian	Northern Bavarian Central Bavarian	0.24 0.24	✗ ✗	✓	◐ Lameli (2013)	✗ ✗
	Low Franconian	Westphalian	0.42	✓	✓	✗	◐
	Upper Saxon Thuringian	Thuringian Upper Saxon	0.20 0.20	◐ ✗	◐	✓ Purschke (2011)	✗ ◐
	East Hessian	Central Hessian	0.24	✓	✓	✗	✗
Low recall	Eastphalian	-	0.29	✗	-	-	✓
	North Hessian	-	0.35	◐	-	-	◐
	Moselle Franconian	-	0.29	✗	-	-	◐

Table 3: Severe confusions (>20%) and low correct classification (<40%). Rate refers to either confusion rate or to correct classification rate. Reasons are evidence indicators: Spk = number of speakers, Geo = geographical proximity, Sim = dialectal similarity, Self = speaker self-assessment.

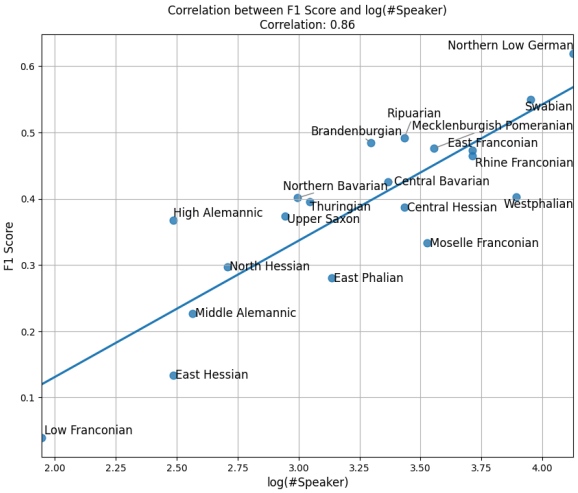


Figure 6: Relationship between the F1-score and the logarithm of the number of speakers per dialect.

vates the focus on the number of speakers rather on the number of samples.

To further characterize this relationship, we estimate a saturation point at approximately 57 speakers, determined by a slope threshold of 0.001. Beyond this point, additional speakers yield only marginal F1-score gain, suggesting diminishing returns of speaker diversity for this model.

Two additional thresholds are derived from the distribution of predicted dialect-wise F1-scores. A threshold at 0.329, corresponding to the 33.33rd percentile, is crossed at approximately 20.3 speakers, while a threshold at 0.2639 (20th percentile) is crossed at approximately 14.6 speakers. In practical terms, dialects represented by fewer than about 20 speakers are likely to fall within the lower third of classification performance, whereas dialects with fewer than roughly 15 speakers are expected to be among the lowest-performing fifth, as for East Hessian and Low Franconian. These thresholds directly motivate the speaker-based categorization used in Table 3.

6.2. Dialectal Similarity

Dialectal similarity provides a plausible explanation for several of the systematic confusions observed in the classification results. In particular, the frequent mutual confusion between Thuringian and Upper Saxon is consistent with findings reported by Purschke (2011, pp. 292f.). Purschke shows that listeners do not exhibit a clear or stable distinction between speakers of Thuringian and Upper Saxon. This pattern suggests the presence of a higher-level regional variety rather than two clearly separable dialects (see Lameli (2013)).

For the dialect pair North Bavarian and Central Bavarian, similarly high confusion rates are observed. In this case, however, the number of speakers differs more substantially (20 speakers for North Bavarian and 29 for Central Bavarian) than for the Thuringian–Upper Saxon pair (19 and 21 speakers, respectively). This imbalance may contribute to the asymmetric confusion pattern, in which North Bavarian is more frequently classified as Central Bavarian than vice versa. Both varieties of Bavarian are grouped within a broader Bavarian dialect region in regional speech classifications (Lameli, 2013, p. 194), which supports the assumption of dialectal similarity.

A comparable situation can be observed for Middle Alemannic and Swabian. According to Lameli (2013), both varieties are situated within the same

larger Alemannic dialect area. However, quantitative distance measures indicate that Swabian shows a considerable degree of structural distinctiveness within this continuum (Lameli, 2014). Consequently, dialect similarity alone is unlikely to be the primary driver of their frequent confusion.

6.3. Speaker Self-Assessment

Figure 7 shows the distribution of self-assessed speaking proficiency across dialects. Self-assessment scores range from 0 ("not at all") to 6 ("perfect") and reflect speakers' perceived active competence in their local dialect.

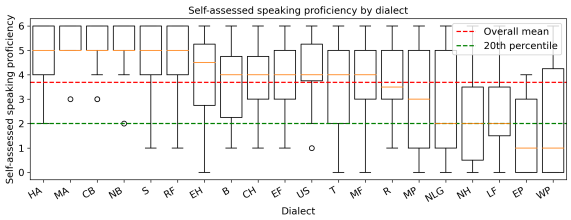


Figure 7: Distribution of self-assessed speaking proficiency by dialect. The red dashed line marks the overall mean, the green dashed line the 20th percentile. Abbreviations as in Figure 4.

A statistically significant correlation is observed between self-assessed speaking proficiency and classification performance (Spearman $\rho = 0.28$, $p < 0.001$). Although this correlation is not strong, its high statistical significance indicates that speakers who perceive themselves as more competent in their dialect tend to be classified more accurately by the model. Importantly, speaker self-assessment does not constitute an objective ground truth for dialect proficiency. Nevertheless, the observed correlation suggests that variation in dialect competence contributes to classification performance and therefore represents a relevant sociolinguistic factor.

The red dashed line in Figure 7 represents the overall mean self-assessment score, while the green dashed line marks the 20th percentile, indicating the threshold below which 20% of all responses fall. This percentile is used as a reference point for identifying dialects with comparatively low self-assessed competence.

For several frequently confused dialect pairs, including Central Bavarian–North Bavarian and Upper Saxon–Thuringian, self-assessment scores are at or above the overall average. This suggests that the observed confusions are unlikely to be primarily driven by low dialect competence, but rather by dialectal similarity. In contrast, Low Franconian, which is often confused with Westphalian, exhibits self-assessment scores slightly above the 20th percentile. While reduced dialect competence among Low Franconian speakers may contribute to confusion in this case, it is unlikely to be the dominant factor. A markedly different pattern is observed for Eastphalian. This dialect shows the lowest self-assessed speaking proficiency among all dialects in the dataset. The consistently low self-assessment scores strongly suggest that limited active dialect competence among speakers is a major contributing factor to the poor classification performance observed for Eastphalian.

6.4. Geographical Proximity

Across all frequently confused dialect pairs in Table 3, the true and predicted dialects are either adjacent or separated only by a transition area.

The only case in which the dialects are not strictly adjacent is the pair Upper Saxon and Thuringian. However, these varieties are separated merely by a transition zone and have already been shown to exhibit a high degree of dialectal similarity (cf. Section 6.2). Consequently, this case does not contradict the general pattern of spatial closeness among confused dialects.

Overall, the consistent spatial proximity of confused dialects suggests that the model captures geographically grounded variation patterns. Neighboring dialects tend to share phonetic and prosodic characteristics, which likely results in similar embedding representations and, in turn, higher confusion rates. Nevertheless, geographical proximity alone is insufficient to explain the observed patterns, as spatial adjacency does not necessarily imply linguistic similarity. The effect of geographical proximity is therefore interpreted as supportive rather than primary and is closely intertwined with dialectal similarity.

6.5. Summary

The analysis reveals that misclassifications are not random, but follow systematic patterns that can largely be explained by a combination of speaker distribution, dialectal similarity, geographical proximity, and speaker self-assessment.

Dialectal similarity emerges as a primary driver of bidirectional confusion. Dialect pairs that are confused in both directions, such as Central Bavarian–North Bavarian and Upper Saxon–Thuringian, are also documented as structurally close in the dialectology literature. These findings indicate that some dialect distinctions may be inherently difficult to model, even with increased amounts of data, and raise the possibility that such varieties could be treated as part of broader dialect regions.

The number of speakers per dialect plays a substantial role in asymmetric confusion patterns. Dialects with fewer speakers are frequently classified

as closely related dialects with larger speaker populations. This effect suggests that speaker diversity strongly influences the robustness of learned dialect representations. However, it remains unclear whether increasing the number of speakers for underrepresented dialects would fully resolve these confusions or whether strong dialectal similarity would continue to limit separability.

Speaker self-assessment shows a statistically significant but comparatively weak association with classification performance. In most confusion cases, speakers rate their dialect competence at or above the overall average, indicating that low self-assessed proficiency is not the primary cause of confusion. An exception is Eastphalian, where consistently low self-assessment scores coincide with particularly poor classification performance, suggesting limited active dialect competence as a key contributing factor.

An important exception is Moselle Franconian, which exhibits a particularly low recognition rate despite a comparatively large number of speakers and average self-assessment scores. In this case none of our tested factors sufficiently explain the observed confusion patterns, suggesting that additional factors may play a role. This is consistent with its characterization as a gradient, non-categorical variety within the Rhenish fan (Herrgen et al., 2019).

Overall, the analysis suggests that dialect classification performance is shaped by an interaction of linguistic similarity and data-related factors rather than by any single variable alone. While increasing speaker diversity may improve performance for some dialects, certain confusions appear to reflect genuine dialectal closeness that may remain challenging to resolve even with additional data.

These findings motivate a closer examination of how the model behaves in dialect transition areas, where linguistic boundaries are inherently gradient.

7. Beyond Classification: Predicting Transition Areas

Beyond discrete dialect classification, the model can be used to analyze dialect transition areas by examining the distribution of predicted dialect labels across geographical locations. Figure 8 visualizes this behavior by aggregating classification results at the level of recording locations. Each dialect is represented by a distinct color, while pie charts illustrate how often audio samples from a given location are assigned to the respective dialects. North Bavarian and Central Bavarian are here combined into a single dialect region, as well as Upper Saxon and Thuringian, based on the findings in Section 6.

To enable a meaningful analysis of transition areas, a confidence threshold was introduced to identify audio segments that cannot be reliably assigned to a single dialect. Segments that do not exceed this threshold are marked as “not properly classified” and are visualized in white. The threshold was selected based on the relationship between the proportion of unclassified segments and speakers’ self-assessed dialect proficiency. Strong negative correlations were observed for threshold values in the range between 0.25 and 0.40, indicating that higher self-assessed proficiency is associated with fewer unclassified segments in this range. A threshold of 0.40 was chosen as a compromise at the upper end of this range, ensuring a relatively strict confidence requirement while preserving the negative association.

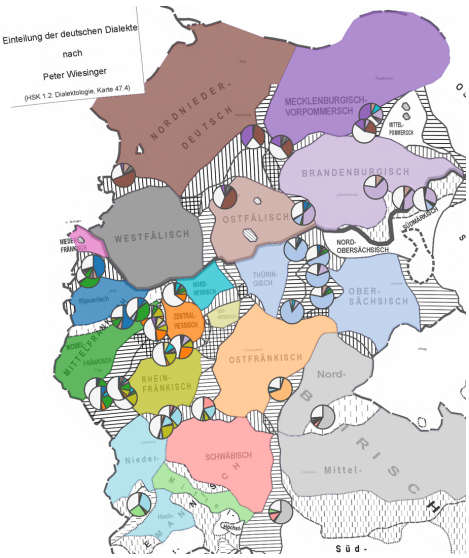


Figure 8: Predicted dialect distributions in transition areas. Colors denote dialects; pie charts show assignment proportions per location. White segments indicate unclassified samples below the confidence threshold.

Across most transition areas, the model predominantly assigns audio samples to the geographically adjacent dialect regions. Locations associated with dialects that already exhibit high confusion rates during training—such as Central Hessian, East Hessian, and Moselle Franconian—show particularly large proportions of unclassified segments. This suggests that the model’s uncertainty in transition areas mirrors the difficulty of distinguishing these dialects more generally.

The transition maps further allow for generational comparisons. Figure 9 compares predictions for the younger and the older generation only. Consistent with the generational analysis in Section 5, the younger generation exhibits a larger proportion of segments that are not properly classified. When comparing the two generations, shifts in dominant dialect assignments can be observed in selected regions, particularly in the southern area shown in

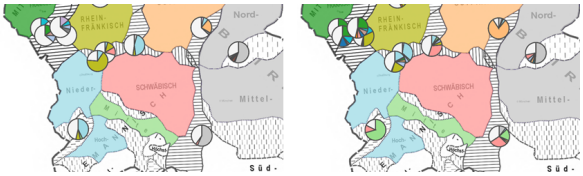


Figure 9: Excerpt from Figure 8 showing selected transition areas. The left panel displays predictions for the younger generation, while the right panel shows predictions for the older generation.

the enlarged excerpt. As illustrative examples, we consider the southern locations in the lower left and the lower right of the maps in both panels (younger and older). At these locations, which are associated with the High- and Middle Alemannic areas on the one hand and the Swabian and Bavarian areas on the other, the predicted dialect proportions differ between the younger and the older generation. This difference can also be interpreted as language change, which has already been noted in other research (Lameli (2013, p. 314)).

Taken together, these results demonstrate that the proposed model is not limited to categorical dialect classification but can also be used to explore gradient dialectal variation and generational change in transition areas.

8. Discussion

The results of this study indicate that low classification performance for certain dialects is not primarily caused by model limitations but reflects fundamental properties of dialect data. As also noted by Jokisch and Dobbriner (2019), closely related dialects can be difficult to distinguish even for human listeners, suggesting that systematic confusions often arise from inherent linguistic similarity rather than insufficient modeling.

More generally, dialects cannot be treated as strictly discrete categories. Speakers do not necessarily realize all features of a single dialect and may simultaneously use features from neighboring varieties. Dialects are therefore better modeled as fuzzy sets rather than hard classes (Ferragne and Pellegrino, 2007). This explains why certain samples remain difficult to classify with high confidence.

Generational analyses show that younger speakers achieve lower classification performance and produce more low-confidence or unclassified segments. At the same time, cross-generational training improves performance for all age groups, indicating that core dialectal features remain shared across generations despite a gradual attenuation among younger speakers.

Speaker diversity is a central aspect of dialect classification. Classification performance correlates strongly with the number of speakers, which in turn is strongly correlated with the number of samples. Nevertheless, speaker diversity is particularly important, as increasing the amount of data from only few speakers risks reinforcing speaker-specific patterns rather than dialectal variation. Unequal speaker distributions thus constitute a structural challenge, consistent with observations by Jokisch and Dobbriner (2019) and findings on diminishing returns reported by Elleuch et al. (2025).

Finally, perfect classification accuracy cannot be expected for dialect data. Not all speech samples are dialectal and assigning speakers to dialects based on place of origin does not guarantee exclusive dialect use. Moreover, as also noted by Jokisch and Dobbriner (2019), direct comparisons across studies remain difficult due to differences in corpora, class definitions, and speaker constellations. The same limitations apply to the present work, and absolute performance values should therefore be interpreted with caution.

9. Conclusion

This work shows that automatic dialect classification should not be treated as a purely accuracy-driven task. Systematic misclassifications and model uncertainty are informative and reflect linguistic structure rather than random model failure. Consequently, dialect classification models should be evaluated not only by how well they separate classes, but by how faithfully their errors reflect linguistic structure.

A detailed error analysis identified four main factors influencing confusion and low performance: speaker diversity, dialectal similarity, geographical proximity, and speaker self-assessment. Among these, the number of speakers per dialect is the most influential model-related factor, while dialectal similarity represents an inherent linguistic limitation. The analyses confirm that closely related dialects lead to systematic confusions and low-confidence predictions, consistent with dialectological accounts of gradient variation.

Generational analyses further show that dialect classification benefits from cross-generational training. Although younger speakers exhibit reduced dialectal distinctiveness, core dialectal features remain shared across generations.

Overall, the proposed model serves not only as a classifier but also as a tool for dialectological analysis: among other applications, it can, as shown in this paper, (i) reveal dialect similarity structure from systematic confusions, (ii) support apparent-time analyses of how dialects change across generations, and (iii) characterize speakers or locations in terms of dialectal affinities (illustrated here using transition areas).

10. Acknowledgements

This research is supported by the Federal Ministry of Research, Technology and Space (BMFTR) (grant AnDy 16DKWN007) and the Academy of Sciences and Literature Mainz (grant REDE 0404), the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence LAMARR22B, and the Research Center Deutscher Sprachatlas Marburg. We are grateful to two anonymous reviewers for helpful comments.

11. Bibliographical References

Sakhia Darjaa, Róbert Sabo, Marián Trnka, Milan Rusko, and Gabriela Múcsková. 2018. Automatic recognition of slovak regional dialects. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 305–308. IEEE.

Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. Adi-20: Arabic dialect identification dataset and models. In *Interspeech 2025*, pages 2775–2779. ISCA.

Emmanuel Ferragne and François Pellegrino. 2007. Automatic dialect identification: A study of british english. In *Speaker Classification II: Selected Projects*, pages 243–257. Springer.

Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. 2025a. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Interspeech 2025*, pages 2780–2784.

Lea Fischbach, Akbar Karimi, Alfred Lameli, and Lucie Flek. 2025b. Edaudio: Easy data augmentation for dialectal audio. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 363—368.

Karzan Ghafoor, Sarkhel Taher, Karwan Hama Rawf, and Ayub Abdulrahman. 2025. The improved Kurdish dialect classification using data augmentation and ANOVA-based feature selection. *ARO - The Scientific Journal of Koya University*, 13:94–103.

Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2024. Continual learning under language shift. In *International Conference on Text, Speech, and Dialogue*, pages 71–84. Springer.

Joachim Herrgen, Jürgen Erich Schmidt, and Robert Möller. 2019. Historisches westdeutsch/rheinisch (moselfränkisch, ripuarisch, südniederfränkisch). In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum. Ein internationales Handbuch der Sprachvariation*, volume 4 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 515–550. De Gruyter Mouton, Berlin/Boston.

Nina Hosseini-Kivanani, Christoph Schommer, and Peter Gilles. 2025. Voices of luxembourg: Tackling dialect diversity in a low-resource setting. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 143–152.

Oliver Jokisch and Johanna Dobbriner. 2019. Textindependent dialect classification in read and spontaneous speech. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 350–354.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*, volume 54. Walter de Gruyter.

Alfred Lameli. 2014. Distanz als raumstrukturelle eigenschaft dialektaler kontaktsituationen. In Dominique Huck, editor, *Alemannische Dialektologie: Dialekte im Kontakt*, Zeitschrift für Dialektologie und Linguistik. Beihefte, pages 67–86, 297–300. Franz Steiner Verlag, Stuttgart. Beiträge der 17. Arbeitstagung für alemannische Dialektologie, Straßburg, 26.–28.10.2011.

Alfred Lameli. 2025. Gesprochenes Deutsch in den Regionen. Eine Standortbestimmung für die Bundesrepublik Deutschland. In Nadine Proske, Thilo Weber, Monika Dannerer, and Arnulf Deppermann, editors, *Gesprochenes Deutsch. Struktur, Variation, Interaktion*, pages 51–79. De Gruyter, Berlin and Boston.

Aynalem Tesfaye Misganaw and Sabine Roller. 2022. German dialect identification and mapping for preservation and recovery. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 65–69.

Hermann Niebaum and Jürgen Macha. 2014. *Einführung in die Dialektologie des Deutschen*. De Gruyter, Berlin, Boston.

Christoph Purschke. 2011. *Regionalsprache und Hörerurteil*. Franz Steiner Verlag.

Jürgen Erich Schmidt and Joachim Herrgen and Roland Kehrein and Alfred Lameli and Hanna Fischer. 2020–. *Regionalsprache.de: Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Philipps-Universität Marburg. Digitale Sprachressource. Bearbeitet von Lisa Dücker, Robert Engsterhold, Marina Frank, Heiko Girnth, Simon Kasper, Juliane Limper, Salome Lipfert, Georg Oberdorfer, Tillmann Pistor, Anna Wolańska. Unter Mitarbeit von Dennis Beitel, Lea Fischbach, Milena Gropp, Heiko Kammers, Maria Luisa Krapp, Vanessa Lang, Salome Lipfert, Nathalie Mederake, Jeffrey Pheiff, Bernd Vielsmeier. Studentische Hilfskräfte.

Antonio Sciarretta. 2024. Dialectometry-based classification of the central–southern italian dialects. *Journal of Linguistic Geography*, 12(1):13–23.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157.

Samuel Stucki and Patrik Randjelovic. 2021. Automatic detection of swiss german dialects using wav2vec. Master's thesis, Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland.

Peter Wiesinger. 1983. Die Einteilung der deutschen Dialekte. In Werner Besch, editor, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, volume 1.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 807–900. Berlin/New York: de Gruyter, Berlin, New York.

12. Language Resource References

Joel Shor and Subhashini Venugopalan. 2022. *TRILLsson: Distilled Universal Paralinguistic Speech Representations [Pre-trained Model]*.

FLEURS-Kobani: Extending the FLEURS Dataset for Northern Kurdish

Daban Q. Jaff[1,2], **Mohammad Mohammadamini**[3]
[1]Erfurt University, Erfurt, Germany
[2]Koya University, Koysinjaq, Iraq
[3]LIUM, Le Mans University, Le Mans, France
daban.hamad_ameen@uni-erfurt.de, mohammad.mohammadamini@univ-lemans.fr

Abstract

FLEURS offers n-way parallel speech for 100+ languages, but Northern Kurdish is not one of them, which limits benchmarking for automatic speech recognition and speech translation tasks in this language. We present FLEURS-Kobani, a Northern Kurdish (ISO 639-3 KMR) spoken extension of the FLEURS benchmark. The FLEURS-Kobani dataset consists of 5,162 validated utterances, totaling 18 hours and 24 minutes. The data were recorded by 31 native speakers. It extends benchmark coverage to an under-resourced Kurdish variety. As baselines, we fine-tuned Whisper v3-large for ASR and E2E S2TT. A two-stage fine-tuning strategy (Common Voice→FLEURS-Kobani) yields the best ASR performance (WER 28.11, CER 9.84 on test). For E2E S2TT (KMR→EN), Whisper achieves 8.68 BLEU on test; we additionally report pivot-derived targets and a cascaded S2TT setup. FLEURS-Kobani provides the first public Northern Kurdish benchmark for evaluation of ASR, S2TT and S2ST tasks. The dataset is publicly released for research use under a CC BY 4.0 license.

Keywords: Few-shot Learning, Speech Translation, Automatic Speech Recognition, Low-resource Languages, Northern Kurdish

1. Introduction

Expanding available speech resources for new languages can significantly reduce the effort required for data curation across multiple tasks. Among widely used multi-task speech resources, the FLEURS (Conneau et al., 2022) dataset supports few-shot learning and serves as a benchmark for several speech technologies, including automatic speech recognition (ASR), speech-to-text translation (S2TT), and Speech-to-Speech Translation (S2ST), across more than 100 languages. FLEURS is a multilingual speech extension of FLORES (Goyal et al., 2021) that provides spoken data for 101 of the 200 languages available in FLORES. Consequently, around 100 languages do not have corresponding speech data in FLEURS. Northern Kurdish is one of these languages. In this paper, we introduce the FLEURS-Kobani dataset, a spoken extension of the FLORES portion for Northern Kurdish.

A major obstacle in low-resource and multilingual NLP is not only training data scarcity, but also the lack of high-quality, standardized evaluation benchmarks that enable fair comparison across languages and research groups. FLORES-101 (Goyal et al., 2021) was introduced to address this problem for machine translation (MT) by providing a carefully constructed benchmark: 3,001 sentences sampled from English Wikipedia and professionally translated into 200 languages under a controlled process, with broad topical coverage (NLLB Team et al., 2022).

Building on this idea, FLEURS (Conneau et al., 2022) extends the FLORES evaluation benchmark by adding the speech modality. FLEURS is an n-way parallel speech benchmark created by recording speech for sentences drawn from FLORES, yielding roughly ≈12 hours of speech per language. FLEURS is designed not as a single-task corpus but as a benchmark dataset for few-shot learning that can be used for multiple speech processing tasks, including automatic speech recognition (ASR), Spoken Language Identification, Speech-to-Speech Translation (S2ST) and Speech-to-Text-Translation (S2TT) tasks (Conneau et al., 2022).

However, while the FLEURS benchmark includes Central Kurdish (ISO 639-3 CKB), it does not include Northern Kurdish, leaving it underrepresented in widely used benchmarking suites. Motivated by this gap and by the benchmark-extension approach exemplified by FLEURS, we introduce FLEURS-Kobani, a Northern Kurdish (KMR) multitask dataset built by recording 2,000 unique sentences from the Northern Kurdish portion of FLORES. The dataset comprises 5,162 utterances totaling 18 hours and 24 minutes of speech. To the best of our knowledge, this dataset is the first attempt to curate speech translation resources for Northern Kurdish.

1.1. Northern Kurdish

Kurdish is an Indo-European language spoken across the Kurdistan region (Iraq, Iran, Turkey, and Syria) as well as by a sizable population in the diaspora. The number of native speakers is commonly reported to be over 30 million (Öpengin, 2021). Its

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 104–109
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

broad geographic dispersion and socio-political history have contributed to strong internal diversity, and Kurdish is often described as a dialect continuum (Eppler and Benedikt, 2017) or a macrolanguage (Sheyholislami, 2015).

Dialectological descriptions commonly distinguish major groupings such as Northern Kurdish (KMR), Central Kurdish (CKB), Southern Kurdish (SDH), Zazaki and Hewrami (Haig and Öpengin, 2014; Mohammadirad, 2026). Kurdish diversity is further reflected in writing practices shaped by geopolitics: Latin-based orthography is widely used for Northern Kurdish and Zazaki, whereas Arabic-based scripts are typical for Central Kurdish, Southern Kurdish and Hewrami groups (Sheyholislami, 2015; Hassani, 2017).

The focus of this study is Northern Kurdish, which is spoken in Turkey, Syria as well as parts of Iraq and Iranian Kurdistan, with additional speaker communities in countries such as Armenia, Lebanon and the east of Iran (Khorasan) (Hassanpour, 1992; Izady, 1992). Although Northern Kurdish is spoken by the majority of Kurdish speakers, it is often restricted or prohibited as a medium of education in Iran, Turkey, and Syria where the majority of speakers are situated (Bourgeois-Gironde et al., 2026).

1.2. Related Work

In recent years, there have been significant efforts to provide speech resources for Kurdish dialects. In (Veisi et al., 2022), a phonetically balanced ASR dataset is designed for Central Kurdish. In (Ahmadi et al., 2024), an ASR dataset for Central Kurdish sub-dialects derived from TV series and comedy programs is developed. Common Voice 18 includes three Kurdish dialects, providing 68 hours of validated Northern Kurdish speech, 134 hours of Central Kurdish, and 1 hour of Zazaki (Ardila et al., 2020). In (Hameed et al., 2025), a collection of ASR resources for low-resource Middle Eastern languages is presented, which includes Laki, Hawrami, and Southern Kurdish branches of Kurdish.

To combat the resource scarcity in the speech translation domain, recent work has explored Kurdish S2TT via pseudo-labeling, using an ASR+MT pipeline to create large pseudo-labeled Kurdish–English data for training end-to-end models (Mohammadamini et al., 2025a). Kuvost is a large-scale English to Central Kurdish S2TT dataset provided by the translation of English common Voice to Kurdish (Mohammadamini et al., 2025b). The FLEURS dataset (Conneau et al., 2022) includes Central Kurdish but not Northern Kurdish.

Taken together, prior Kurdish speech research is unevenly distributed across dialects and tasks. The most substantial ASR and S2TT resources are concentrated on Central Kurdish. While Northern Kurdish is occasionally included in multilingual asr efforts (Omnilingual ASR team et al., 2025), it remains largely missing from standard evaluation benchmarks for ASR, S2TT and S2ST tasks.

2. Data Recording and Validation

2.1. Recording

Data were recorded over a two-month period, from late April 2025 to late June 2025 at the University of Kobani via the Jiridastkrd[1] platform. A total of 31 speakers participated in the recordings: consisting of 26 female and 5 male speakers. Data collection followed three main phases: recruitment, outreach, and training.

Recruitment and team setup: In mid-April 2025, third-year students from the English Department, Faculty of Education were introduced to the project and invited to a two-hour orientation on procedures for recording Northern Kurdish sentences using the Jiridastkrd.com platform. Participants were informed that participation was voluntary, that they could withdraw at any time, and that they could request deletion of their recordings without needing to provide a reason. Of the nine students initially contacted, eight continued with the project.

Outreach and speaker diversification: Because the initial student group was small and entirely female, the eight participants were asked to help recruit additional speakers, especially male speakers, to improve gender and speaker coverage.

Training and recording procedure: Students were trained on (i) reading the prompt before recording, (ii) how to begin reading only after pressing the record button, and (iii) how to upload recordings via the platform. Additional recruits received the same orientation.

2.2. Validation

Validation was conducted within the Jiridastkrd.com workflow by reviewing each recording together with its associated Northern Kurdish prompt and metadata. We applied two classes of checks: prompt conformity (whether the speaker read the intended sentence) and audio quality (whether the recording is sufficiently intelligible). Recordings were flagged for exclusion if they met any of the following criteria:

(i) **Prompt mismatch/misreading:** the spoken content does not match the assigned Northern Kurdish prompt (e.g., reading a different sentence or substantial divergence from the prompt).

[1] jiridastkrd.com

(ii) **Truncation:** a word (or multiple words) is cut off at the beginning, middle, or end of the recording.

(iii) **Insertions:** the speaker adds extra word(s) not present in the prompt (beyond minor filled pauses).

(iv) **Pronunciation errors:** the pronunciation renders a word difficult to recognize or changes the intended lexical item; as exceptions, we did not enforce a single "correct" reading for foreign proper names and abbreviations, which appear in the FLEURS prompts and were read inconsistently across speakers.

(v) **Excessive background noise:** or overlapping speech that reduces intelligibility.

(vi) **Very low recording level:** (low volume) such that the speech cannot be heard clearly under normal listening conditions.

Socio-political and technical constraints in the Rojava region directly influenced the recording process. A primary limitation is the significant demographic imbalance; the project was initiated with a core team of eight female students, and despite subsequent outreach efforts, the team remains heavily skewed, with 26 female speakers compared with only 5 male participants. Data collection was further hampered by unstable internet connectivity and limited digital infrastructure outside of the University of Kobani, and the reliance on personal cellphones for recordings contributed to a high rate of invalid entries due to varying microphone quality and technical inconsistencies.

Furthermore, many participants reported limited formal education in Kurdish, with most having started studying it only in seventh grade as an elective subject. This negatively affected reading fluency, increased variability in the recordings, and resulted in specific linguistic inconsistencies, particularly in the pronunciation of abbreviations and foreign proper names, including the names of cities and countries. Although direct prompt mismatches were relatively limited, other technical errors were very common, including truncation in which words were cut off, excessive background noise, and pronunciation errors that rendered words difficult to recognize. Additionally, a significant number of recordings were excluded due to very low recording levels that made speech unintelligible. As a result, of the 7,897 Northern Kurdish recordings, 5,162 were marked as good quality (65.5%), while 2,735 were marked as low quality (34.5%).

3. Data Specifications

The resulting validated data comprise 5,162 utterances, equivalent to 18 hours and 24 minutes of speech. The validated utterances are partitioned into three subsets, according to the English version of the FLEURS dataset. For a given file, if its English translation matches a sentence in the English FLEURS dataset, it is assigned to the corresponding train/dev/test split. We found 149 samples that did not match any sentence in the English FLEURS dataset. These utterances are placed into the dev set in order to increase the number of files in this partition from 342 to 491 utterances. The numbers of utterances and the duration of each partition are presented in Table 1. The dataset can be accessed through a dedicated Hugging Face repository [2].

Partition	Train	Dev	Test
Utterances	3,870	491	801
Duration	13:47	1:42	2:55
Speakers	30	20	26

Table 1: FLEURS-Kobani benchmark specifications

4. Experiments and Results

We evaluated our dataset using the Whisper v3-large model for both ASR and S2TT tasks (Radford et al., 2022). The Whisper model was fine-tuned for five epochs with a learning rate initialized at 1e-5. Evaluation on the validation set was performed every 50 training steps. The batch size was set to 4 during training. The checkpoint achieving the best performance on the validation set was selected for evaluation on the test set.

4.1. ASR Results

The dataset was evaluated on the ASR task under three experimental conditions. First, we fine-tuned the Whisper model using the Northern Kurdish subset of Common Voice, which contains 68 hours of validated speech. This configuration achieved a WER of 36.28 on the development set and 37.02 on the test set. Second, we fine-tuned the Whisper v3-large model using the training portion of FLEURS-Kobani, obtaining a WER of 34.86 on the development set and 38.09 on the test set. Finally, we fine-tuned the Whisper model on FLEURS-Kobani after an initial fine-tuning stage on Common Voice. This two-stage fine-tuning strategy resulted in a WER of 28.53 on the development set and 28.11 on the test set. The results are reported in Table 2.

4.2. E2E S2TT Results

As a first E2E S2TT experiment, we fine-tuned the Whisper v3-large model on the training portion of

[2] https://huggingface.co/datasets/aranemini/northern-kurdish-fleurs

-	CV	F-K	CV → F-K
DEV	12.33/36.28	13.98/34.86	10.11/28.53
TEST	12.28/37.02	15.97/38.09	9.84/28.11

Table 2: ASR results on KMR FLEURS benchmark (CER/WER), CV: Common Voice, F-K: FLEURS-Kobani

the FLEURS-Kobani dataset. This configuration achieved BLEU scores of 9.16 and 8.68 on the development and test sets, respectively. For other languages available in the FLEURS dataset, English was used as a pivot language. Using this pivot, we created parallel KMR–FR, KMR–NL, KMR–CKB, and KMR–FA versions of the train/dev/test splits of the FLEURS-Kobani dataset. We observed that target languages such as NL and FR, which are closer to the pivot language, yield better results. For the KMR–CKB and KMR–FA pairs, despite their linguistic similarity to KMR, we did not achieve significant results. This behavior may be attributed to divergences introduced by lexical and typological differences between the target languages and the pivot language. The E2E S2TT results are presented in Table 3.

Target	DEV	TEST
EN	9.16/34.72	8.68/34.15
FR	5.00/27.94	6.16/29.78
NL	2.70/25.71	2.82/26.62
CKB	0.84/19.11	1.24/20.83
FA	0.94/20.15	1.36/21.05

Table 3: E2E S2TT results (BLEU/ChrF++) from Northern Kurdish (KMR) to five target languages/dialects.

4.3. Cascaded S2TT

In the final experiment, we evaluated our dataset on the cascaded S2TT task. First, we transcribed the KMR audio using an ASR model (Whisper fine-tuned on CV and FLEURS-Kobani). Then, the ASR outputs were translated using NLLB-1.3B (NLLB Team et al., 2022). The cascaded results are reported in the ASR+MT column of Table 4. The first column of Table 4 presents the machine translation (MT) results obtained by translating the references, which serve as the baseline.

Split	MT (Base)	ASR+MT
Dev	24.98/51.68	16.98/43.91
Test	27.36/52.96	19.60/45.51

Table 4: Cascaded (KMR→EN) S2TT results (BLEU/ChrF++): MT(Base) translation of references using NLLB; ASR+MT translates ASR outputs using the NLLB model.

5. Conclusion

This paper introduced FLEURS-Kobani, a speech extension of FLEURS for Northern Kurdish (Kurmanji; KMR). The dataset helps address a key resource gap by enabling benchmarking of core speech technologies such as ASR, speech-to-text translation (S2TT), and Speech-to-Speech Translation (S2ST) for Northern Kurdish. By releasing FLEURS-Kobani along with baseline results, our aim is to lower the barrier for future research and accelerate progress on Kurdish speech technology, with benefits for accessibility and language inclusion. Since there are two variables—fluency and dialect—in the utterances, a more granular evaluation at the speaker level, as well as an analysis of regional dialect variability in comparison to the written form, will provide a clearer view of the obtained results in future research. This work encourages future research to expand coverage to additional Kurdish varieties.

Acknowledgements

The authors express their deepest gratitude to the following individuals for their dedicated support and contributions to the project:

Arîvan Misî, **Bahoz** Makû, **Bushra** Ali, **Cedra** Hami, **Diane** Mohamed, **Dilava** El Ehmed, **Dilê** Berkel, **Eslem** Ahmed, **Falak** Hitto, **Hacra** Îsa, **Kaniwar** Hisên, **Kemal** Besrawî, **Mehmûd** Elî, **Mehrîvan** Hussain, **Mistefa** Temo, **Narîn** Bedir, **Nazdar** Ebde, **Naze** Afandi, **Nujeen** Ebda, **Ranya** Heyder, **Rojan** Bozan, **Roz** Mistefa, **Ruho** Mihemed, **Salan** Bozan, **Sara** Hussain, **Serbest** Bozan, **Shilan** Afandi, **Silva** Haj Ali, **Talîn** Mistefa, **Yasmin** Mustefa, **Zehra** Xelîl.

Our sincere thanks also go to **Kobani University** for facilitating the project and providing the necessary institutional support. Additionally, we would like to extend a special note of thanks to **Razaw Bor** for her invaluable assistance in validating portions of the dataset.

Daban Q. Jaff gratefully acknowledges the support of the **Deutscher Akademischer Austauschdienst (DAAD)** through a PhD research grant (Grant No. 57645448) for his doctoral studies at the University of Erfurt (Host: **Language and Its Structure, Prof. Dr. Beate Hampe**). Mohammad Mohammadamini acknowledges the **LIUM (Le Mans University)** for hosting and providing the computational resources used in the experiments. Finally, the authors thank the anonymous reviewers for their valuable comments.

6. References

Sina Ahmadi, Daban Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. Language and speech technology for Central Kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045, Torino, Italia. ELRA and ICCL.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Sacha Bourgeois-Gironde, Victor Ginsburgh, Hossein Hassani, and Shlomo Weber. 2026. A lingua franca for kurdish populations. In Shlomo Weber and Victor Ginsburgh, editors, *The Palgrave Handbook of Economics and Language*, 2nd edition, pages 575–593. Palgrave Macmillan, Cham.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Eva Eppler and Josef Benedikt. 2017. A perceptual dialectological approach to linguistic variation and spatial analysis of kurdish varieties. *Journal of Linguistic Geography*, 5(2):109–130.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Geoffrey Haig and Ergin Öpengin. 2014. Introduction to special issue - kurdish: A critical research overview. *Kurdish Studies*, 2(2):99–122.

Razhan Hameed, Sina Ahmadi, Hanah Hadi, and Rico Sennrich. 2025. Automatic speech recognition for low-resourced middle eastern languages. In *Interspeech 2025*, pages 733–737. ISCA.

Hossein Hassani. 2017. Kurdish interdialect machine translation. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 63–72, Valencia, Spain. Association for Computational Linguistics.

Amir Hassanpour. 1992. *Nationalism and Language in Kurdistan, 1918–1985*. Mellen Research University Press, San Francisco.

Mehrdad R. Izady. 1992. *The Kurds: A Concise Handbook*. Crane Russak, Washington, DC.

Mohammad Mohammadamini, Daban Jaff, Sara Jamal, Ibrahim Ahmed, Hawkar Omar, Darya Sabr, Marie Tahon, and Antoine Laurent. 2025a. Kuvost: A large-scale human-annotated English to Central Kurdish speech translation dataset driven from English common voice. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 106–109, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, and Antoine Laurent. 2025b. Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of kurdish language). In *Interspeech 2025*, pages 898–902.

Masoud Mohammadirad. 2026. *A Grammar of Hewramî*. Language Science Press, Berlin.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean

Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Jaffer Sheyholislami. 2015. Language varieties of the kurds. In Wolfgang Taucher, Mathias Vogl, and Peter Webinger, editors, *The Kurds: History, Religion, Language, Politics*, pages 30–51. Austrian Federal Ministry of the Interior, Vienna, Austria.

Hadi Veisi, Hawre Hosseini, Mohammad Mohammadamini, Wirya Fathy, and Aso Mahmudi. 2022. Jira: A central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. *Language Resources and Evaluation*, 56(3):917–941.

Ergin Öpengin. 2021. The history of kurdish and the development of literary kurmanji. In Hamit Bozarslan, Cengiz Gunes, and Veli Yadirgi, editors, *The Cambridge History of the Kurds*, pages 603–632. Cambridge University Press.

Exploring the Reusability of Northern Kurdish Resources for Badini Speech Recognition

Mohammad Mohammadamini[1], **Aveen Jalal Mohammed**[2]
Barzan Hussein Mohammed[3], **Dezheen H. Abdulazeez**[2]
Imad Saeed Sadeeq[3], **Dilgash Mohammed Salih Tayib**[3]
Amera Ismail Melhum[2], **Abuobaida Abdullah Dheyab**[3]
[1] LIUM, Le Mans Université, France
[2] Department of Computer, University of Duhok, Kurdistan, Iraq
[3] Department of Kurdish, College of Languages, University of Duhok, Kurdistan, Iraq
first.last@univ-lemans.fr, first.last@uod.ac

Abstract

Badini is a variant of the Kurdish language spoken in the Duhok province of the Kurdistan Region of Iraq. It is written mainly in a modified version of the Arabic script. Although it shares the same script as Central Kurdish (CKB), it is linguistically classified under the Northern Kurdish (KMR) branch. In this paper, we explore the potential and limitations of Northern Kurdish ASR resources for the Badini variant. Firstly, we transliterate the Common Voice 18 dataset from the Latin script into the modified Arabic script and revise it to align with the orthographic conventions of the Badini variant. Additionally, we introduce the first text collection for the Badini variant, containing 14.22 million tokens, which serves as a source for speech synthesis. The third resource developed in this research is a standard speech recognition benchmark recorded by 5 speakers which includes 2 hours and 46 minutes of multi-domain read speech. Results show that combining transliterated and synthetic data significantly improves recognition accuracy, achieving a 6.8% CER and 34% WER. All three resources curated during this research will be made available under the CC BY-NC-ND 4.0 license.

Keywords: Northern Kurdish, Badini Variant, Speech Recognition, Common Voice, Low-resource ASR

1. Introduction

Kurdish population with more than 35 million people speak a multi-dialect continuum categorized into several groups: Central Kurdish (CKB), Northern Kurdish (KMR), Southern Kurdish (SDH) (Matras, 2019), and Hewrami-Zazaki (Todd, 1985). Despite its wide usage, the Kurdish language is considered a low-resource language (Veisi et al., 2020). Its use across four countries, the coexistence of both Arabic and Latin writing systems, and the wide range of dialects have made the development of an inclusive Kurdish automatic speech recognition (ASR) system challenging.

In recent years, there have been significant efforts to develop speech recognition resources for Kurdish dialects. Among the most prominent of these resources, we can point to Common Voice (Ardila et al., 2020), which in its 22nd version includes 136 hours of Central Kurdish, 71 hours of Northern Kurdish, and 2 hours of validated Zazaki speech. [1] Other notable efforts include the Asosoft Speech Corpus (Veisi et al., 2022), which in its first version introduced 46 hours of designed and phonetically balanced audio for Central Kurdish and has been used in several studies with remarkable results (Veisi et al., 2022; Mohammadamini et al., 2025). In other significant endeavors to provide speech recognition resources, (Hameed et al., 2025) has provided ASR resource for Hawrami and Southern Kurdish. FLEURS is another resource that includes Central Kurdish for speech recognition and translation (Conneau et al., 2023).

The scope of the current research is the Badini variant of Northern Kurdish, for which, to the best of our knowledge, there are no specific studies or ASR resources to date. Northern Kurdish is mainly written in the Latin script in Turkey and Syria, while the Badini variant from this group is primarily written in the Arabic script in the Kurdistan Region of Iraq.

In the current research, we explore the reusability of ASR resources of Northern Kurdish in the Badini variant. Since Badini is written in the same script as Central Kurdish, we firstly show the limitations of Central Kurdish ASR for Badini. In fact, Badini is a subgroup of Northern Kurdish but the codified variant of Northern Kurdish is mainly written in the Latin script. In order to show the potential of reusing Northern Kurdish resources for this variant, we transliterated Common Voice of Northern Kurdish into Arabic script and revised it manually to adapt it with the orthographic conventions of Kurdish in the Arabic script (Team, 2025). After demonstrating the potential and limitations of speech corpora from these dialects for the Badini dialect, we propose a text-only data augmentation approach using Badini text and a Northern Kurdish TTS system (Pratap et al., 2024). The synthetic speech is generated from a text corpus which is collected during the current research. The curated Badini text collection includes 14.22 million space

[1] https://commonvoice.mozilla.org/en/datasets

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 110–115
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

separated tokens mainly collected from websites. The third effort in this study is developing a multi-domain ASR benchmark for Badini dialect which is used to evaluate the developed systems. The developed benchmark includes 2 hours and 46 minutes of read speech validated manually.

In the following, in Section 2 we give a brief description of Badini dialect, in Section 3 the proposed methodology of reusing the available resources is presented, in Section 4 the curated resources are described and Section 5 presents the obtained results.

2. Kurdish language and Badini variant

The Kurdish language is an Indo-European language spoken by an estimated 30 to 46 million native speakers across the Kurdistan regions of Iran, Iraq, Turkey, and Syria (Sheyholislami, 2015). In Turkey and Syria, the Kurdish dialects are mainly written in the Latin script, while in Iran and Iraqi Kurdistan, almost all dialects use a modified version of the Arabic script. The scope of our study is Badini, which is a variant of Northern Kurdish. While the majority of Northern Kurdish speakers live in Turkey and Syria and use the Latin script, a significant number of Northern Kurdish speakers live in Iraqi Kurdistan. The spoken Northern Kurdish sub-dialect in these regions is called *Badini*. More precise linguistic classification studies place Badini in the *Southeastern* subgroup of Northern Kurdish, which is spoken in the Hakkari province in Turkey and the Duhok province of Iraqi Kurdistan (Öpengin and Haig, 2014) (Haig, 2018) (Figure. 1).

Both the Latin and Arabic scripts used for Kurdish dialects are almost phonemic (Veisi et al., 2020); there is almost one-to-one correspondence between letters and phonemes. However, this rule is relative. In Table 2, the list of letters in the Badini variant and the Latin script used for Northern Kurdish are presented. As shown in the table, there are six letters (ح، ع، خ، غ، ڵ، ێ) in the Arabic script that are not represented in the Latin script. All of these six letters correspond to phonemes in the Badini variant, but not all of them occur in every subgroup of Northern Kurdish (Öpengin and Haig, 2014), and therefore they are not consistently represented in the script. The /ə/ phoneme is the only phoneme represented in the Latin script that does not have a corresponding letter in the Arabic script. We will return back to this mismatch between the phonemes and its impact on the performance of ASR systems developed for standard variant of Northern Kurdish on Badini variant (see Section 5).

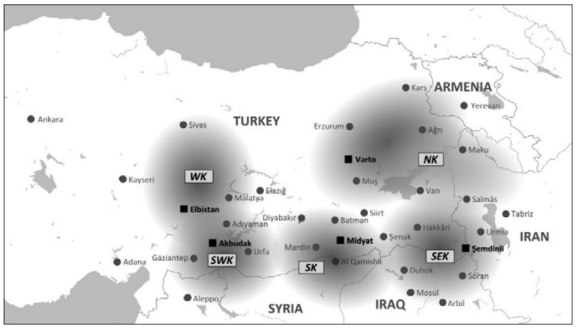


Figure 1: The geography of Badini shown by SEK label as a part of the Southeastern subgroup of Northern Kurdish (Öpengin and Haig, 2014).

Badini Letter			KMR	IPA
Init	Med	Fin		
بـ	ـبـ	ـب	B b	b
د	ـد	ـد	D d	d
جـ	ـجـ	ـج	C c	ʤ
گـ	ـگـ	ـگ	G g	g
ڤـ	ـڤـ	ـڤ	V v	v
ز	ـز	ـز	Z z	z
ژ	ـژ	ـژ	J j	ʒ
غـ	ـغـ	ـغ	–	ɣ
عـ	ـعـ	ـع	–	ʕ
تـ	ـتـ	ـت	T t	t
چـ	ـچـ	ـچ	Ç ç	ʧ
کـ	ـکـ	ـک	K k	k
پـ	ـپـ	ـپ	P p	p
قـ	ـقـ	ـق	–	q
ئـ	ـئـ	ـئ	–	ʔ
هـ	ـهـ	ـھ	H h	h
سـ	ـسـ	ـس	S s	s
شـ	ـشـ	ـش	Ş ş	ʃ
فـ	ـفـ	ـف	F f	f
خـ	ـخـ	ـخ	X x	x
حـ	ـحـ	ـح	–	ħ
ر	ـر	ـر	R r	r
ڕ	ـڕ	ـڕ	RR rr	r
لـ	ـلـ	ـل	L l	l
ڵـ	ـڵـ	ـڵ	–	ɬ
مـ	ـمـ	ـم	M m	m
نـ	ـنـ	ـن	N n	n
یـ	ـیـ	ـی	Y y	j
و	ـو	ـو	W w	w
ی	ـی	ـی	Î î	iː
ا	ـا	ـا	A a	aː
ێـ	ـێـ	ـێ	Ê ê	ɛ
ۆ	ـۆ	ـۆ	O o	o
و	ـو	ـو	U u	u
ووـ	ـووـ	ـوو	Û û	uː
ه	ـه	ـه	E e	æ
-	-	-	I i	ə

Table 1: List of Badini alphabet with initial, medial, and final Arabic forms, KMR equivalents, and IPA.

3. Methodology

For many multi-dialect languages, ASR systems are developed for one or a few standardized dialects, which tend to perform poorly on local variants (Blaschke et al., 2025). In our proposed approach, we explore the reusability of Northern Kurdish resources for Badini, which is a subgroup of this language. Our approach consists of two main parts. First, we use the Common Voice corpus of Northern Kurdish. The scripts are transliterated

using the AsoSoft library[2] into the Arabic script (Mahmudi and Veisi, 2021). In the Arabic script, there is a general tendency to write words together, whereas in the Latin script, words are usually written separately. This simple transliteration therefore causes a significant mismatch between the two scripts. After transliteration, the Arabic script is revised manually to adapt it to the standards commonly accepted in the Arabic orthography. For example, in the Arabic script, the verb "to be" is usually attached to the preceding word, and the second part of circumpositions is often joined with the previous word. During the revision, phonemes and words were not modified; only the rules regarding word joining and separation were adjusted.

In order to compensate for the dialectical differences between standard variation of Northern Kurdish and Badini variant, we use a text-only data augmentation by applying a TTS model available as a part of MMS (Pratap et al., 2024). The flowchart of the proposed approach is shown in Figure 2.

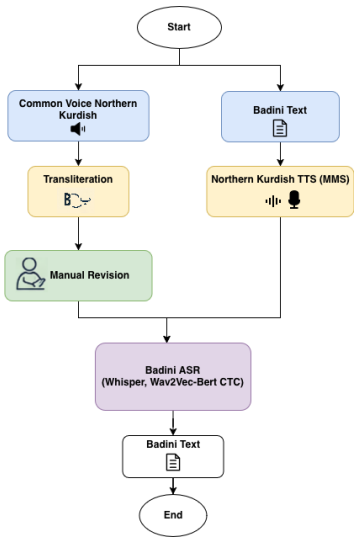


Figure 2: Reusbality of KMR tools and resources for Badini variant

Having the transliterated and synthesised speech we are using two state-of-the art models for Badini speech recognition. The first model is Whisper V3 Large model (Radford et al., 2022) fine-tuned for Badini and the second model is Whav2vec2.0-BERT model optimizes a CTC loss function (Chung et al., 2021). The Whisper model was fine-tuned for five epochs with a learning rate initialized at 1e5.

[2]https://github.com/AsoSoft/AsoSoft-Library

4. Dataset Description

4.1. Transliterated Common Voice

The first dataset used for training the models is the Common Voice 18 transliterated into Arabic script of Kurdish language [3]. This dataset includes 16,808 unique sentences, having several recordings for each sentence the total duration of the validated recorded audio in the transliterated common voice results in 68 hours of speech (Table 3).

4.2. Raw text corpora

During the current study we collected the first text collection for Badini variant from different resources such as news agencies, university websites etc. Using crawling and scraping we collected 70k documents from 15 websites which is equivalent to 14.22 million tokens[4]. The details of the collected text is shown in Table 2

Website Name	Tokens (M)	Documents
Gav	4.56	30964
Speda	0.99	5242
Bas news	1.01	6731
University of Duhok	0.36	2958
University of Zakho	0.06	478
University of Cihan	1.20	9389
Badinan	0.85	1099
Matin journal	1.28	608
Xani agency	0.41	3450
Badinpedia	1.31	1744
kurdax	0,51	1336
kurdislamic	0,54	2421
mzgaft	0,21	343
nrtTv	0,1	542
zebadinan	0,83	2642
Total	14.22	70k

Table 2: Badini Text collection dataset

4.3. Synthetic speech

From the text collection (Table 2), around 119k sentences containing between 3 and 20 tokens were extracted. The extracted sentences were synthesized using the MMS model for Northern Kurdish (Pratap et al., 2024). Among the preprocessing steps, we applied Unicode and number unification (Mahmudi et al., 2019), as well as a specific date and number-to-word conversion for Badini, implemented specifically for this purpose. The synthesized speech from this dataset is 207 hours (Table 3).

4.4. Evaluation Benchmark

The third resource developed during this study is an evaluation benchmark designed for Badini ASR.

[3]https://huggingface.co/datasets/BadiniSpeechNLP/commonvoicebadini

[4]https://huggingface.co/datasets/BadiniSpeechNLP/badinitextcorpora

Model	Utterances	Duration
Common Voice KMR	61.60k	68h
Synthetic data	119k	207h

Table 3: Train speech datasets

The benchmark includes sentences from eight topics. We aimed to create a relatively realistic evaluation set that reflects the main challenges faced by ASR systems, such as code-switching. The scientific category is mainly adapted from textbooks and contains many English scientific terms, while the religion category primarily includes Arabic loanwords. In the evaluation benchmark, we have 623 unique sentences recorded by 5 speakers (2 male and 3 female). The total number of recordings is 1,028 files, resulting in 2 hours and 46 minutes of speech [5]. All recordings in this benchmark were manually validated (Table 4).

Topic	#Samples	Duration	Tokens
art	112	21.48	2173
economy	150	27.70	2944
health	87	14.20	1502
humanities	94	9.86	863
literature	131	10.40	891
politics	48	9.19	998
religion	181	25.42	2945
science	225	48.35	5277
overall	1028	2h46mins	17593

Table 4: Badini ASR evaluation benchmark

5. Experiments and Results

As the first experiment, we evaluated a Whisper-large model fine-tuned on Central Kurdish Common Voice 18 (CV-CKB), achieving 28.22 CER and 76.65 WER, which reflects the high dialectical difference between Central Kurdish and Badini (Table 5). In the second experiment, we fine-tuned both Whisper and Wav2Vec-BERT using the transliterated Northern Kurdish Common Voice (CV-KMR) dataset, where we observed an 11.33 CER with Wav2Vec-BERT. These results show that, although KMR data can lead to significantly better performance than CKB resources due to linguistic similarity, we are still relatively far from a practical ASR model. In the third experiment, we used synthetic Badini speech, obtaining a 9.25 CER. Our results indicate that data generated by TTS models can generalize better than real transliterated Northern Kurdish Common Voice data. In the final experiment, we combined both transliterated Common Voice and synthetic data, achieving a 6.80 CER and 34 WER. Leveraging the potential of both approaches, we achieved a 39% CER relative improvement compared to using only real Northern Kurdish speech.

[5] https://huggingface.co/datasets/BadiniSpeechNLP/badini-asr-benchmark

One reason behind this improvement of using the Badini synthetic data is the compensation of existing lexical difference between standard Northern Kurdish and Badini variant.

Training	Whisper		Wav2Vec-BERT	
	CER	WER	CER	WER
CV-CKB	28.22	76.65	-	-
CV-KMR	24.61	64.30	11.33	50.01
TTS-Badini	18.85	58.46	9.25	39.11
TTS-Bad-CV-KMR	9.00	39.38	6.80	34.41

Table 5: CER and WER on the Badini ASR benchmark for models trained on Central Kurdish, Northern Kurdish, and synthesized Badini data

Based on the results obtained by Wav2Vec-BERT, we present the detailed results for each topic in Table 6. The best performance was achieved in the politics domain, which may be attributed to the fact that a large portion of the synthesized speech originates from this domain. In contrast, the scientific domain showed the poorest performance, likely due to the presence of more challenging sentences containing a higher proportion of non-Kurdish words.

Topic	Whisper		Wav2Vec-BERT	
	CER	WER	CER	WER
art-culture	9.13	42.29	6.90	35.85
economy	8.02	32.03	5.59	30.29
health	9.13	38.56	7.72	35.11
humanities	9.13	42.06	7.81	37.43
literature	9.02	37.60	6.80	32.21
politics	6.40	31.99	5.29	30.89
religion	9.43	38.27	7.07	32.97
science	9.80	44.38	7.21	37.25
overall	9.00	39.38	6.80	34.41

Table 6: CER and WER per domain for Whisper and Wav2Vec-BERT trained on transliterated Common Voice and synthetic speech

One reason behind the low performance of using standard Northern Kurdish ASR data for Badini stems from the mismatch between the two alphabet (shown in Table 2) which cause a significant information loss during transliteration. We analyzed the percentage of these letters in the reference transcripts and predictions produced by the best model (Table 7). In all cases, the percentage of these letters in the predictions was two to three times lower than in the references. This indicates that a main limitation of our proposed approach stems from this issue. In order to have deeper explanation on this type of errors, we extracted the percentage of each letter from the evaluation benchmark. We observed that the frequency of four letters without having a corresponding letter in the Arabic script is very low. Since, the overall occurrence of these letters in the test set is less than one percent, the lexical and morphosyntactic differences (Öpengin and Haig, 2014) between Badini variant and standard Northern Kurdish are the second reason behind

the limited potential of reusing Northern Kurdish resources for Badini variant.

Characters	Reference	Prediction
ح	0.21%	0.07%
ع	0.13%	0.06%
غ	0.09%	0.03%
ڵ	0.16%	0.04%

Table 7: Information loss due to character mismatch between Latin and Arabic script

6. Conclusion

In this paper, we explored the reusability of standard Northern Kurdish ASR resources for the Badini variant. During this study, we prepared three resources: a transliterated and revised version of the Northern Kurdish Common Voice dataset in Arabic script, the first text collection for the Badini variant containing 14.22 million tokens, and a multi-domain ASR benchmark. In the proposed pipeline, which includes transliteration and data augmentation, we demonstrated how existing resources can be reused to improve ASR performance on local variants. However, our results also highlight the limitations of such resources. According to this study, it is necessary to develop dedicated resources for Badini dialect ASR, which will be the focus of our future work.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Mike Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4211–4215.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, pages 913–917.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Geoffrey Haig. 2018. 2.3. northern kurdish (kurmanjî). *The languages and linguistics of Western Asia: An areal perspective*, 6:106.

Razhan Hameed, Sina Ahmadi, Hanah Hadi, and Rico Sennrich. 2025. Automatic Speech Recognition for Low-Resourced Middle Eastern Languages. In *Interspeech 2025*, pages 733–737.

A Mahmudi, H Veisi, M MohammadAmini, and H Hosseini. 2019. Automated kurdish text normalization. In *The second international conference on kurdish and persian languages and literature*.

Aso Mahmudi and Hadi Veisi. 2021. Automated grapheme-to-phoneme conversion for central kurdish based on optimality theory. *Computer Speech Language*, 70:101222.

Yaron Matras. 2019. Revisiting Kurdish dialect geography: Findings from the Manchester database. In *Current Issues in Kurdish Linguistics*, pages 225–241. University of Bamberg Press, Bamberg.

Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, and Antoine Laurent. 2025. Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language). In *Interspeech 2025*, pages 898–902.

Ergin Öpengin and Geoffrey Haig. 2014. Regional variation in kurmanji: A preliminary classification of dialects. In *Kurdish Studies Archive*, pages 171–212. Brill.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25:1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Whisper: Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Jaffer Sheyholislami. 2015. *The Kurds: History, Religion, Language, Politics*, chapter Language Varieties of the Kurds. Austrian Federal Ministry of the Interior.

Badini Orthography Team. 2025. Renusi Badini (In Kurdish). University of Duhok.

Terry Lynn Todd. 1985. *A Grammar of Dimili, also known as Zaza*. Ph.d. dissertation, University of Michigan, Ann Arbor, MI. Near Eastern Studies: Languages and Literatures.

H. Veisi, H. Hosseini, M. MohammadAmini, et al. 2022. Jira: a central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. *Language Resources and Evaluation*, 56:917–941.

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2020. Toward kurdish language processing: Experiments in collecting and processing the asosoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.

Wancho Dialectometry: Community-created data and the Living Dictionaries project

Kellen Parker van Dam
Universität Passau
Passau, Germany
kellenparker.vandam@uni-passau.de

Abstract

Community-created lexical resources for under-documented languages represent an underexplored data source for computational dialectology. This study evaluates the viability of such data for dialectometric analysis, using the Wancho LivingDictionaries project as a case study. Wancho is a Tibeto-Burman language of the Southwestern Patkaian branch, spoken primarily in Longding District, Arunachal Pradesh, India. The dictionary is notable for being entirely community-built and speaker-facing, and uniquely among resources for Northeast India, it incorporates dialect-specific forms spanning village-level geolects and clanlects. We extract dialectal data via automated web scraping and apply a series of preprocessing steps to address inconsistencies in transcription, language labelling, and concept assignment. Pairwise linguistic distances are then computed using Sound Class Alignment (SCA, List 2010), which captures phonological similarity more accurately than raw edit distance by incorporating articulatory feature structure. The resulting distance matrix is analysed through UPGMA hierarchical clustering and NeighborNet split network inference. Despite the dataset's uneven dialect coverage and absence of systematic cognate coding, SCA-based distances recover the traditional Upper / Lower / Middle Wancho distinction and correctly situate transitional varieties. These results hold even for dialects with as few as a dozen attested forms. We show that unlike Bayesian phylogenetic inference which is poorly suited to data of this density and distribution, SCA proves to be a reliable metric. Our findings suggest that SCA distance is robust to the kinds of noise and sparsity characteristic of community-generated lexical data, and that such resources constitute a viable, if imperfect, input for automated dialectometric workflows — particularly in contexts where fieldwork-based data collection is not currently feasible.

Keywords: dialectometry, citizen science, Wancho, Tibeto-Burman

1. Introduction

This study investigates the usability of community-created dialectal data of an under-documented language in order to determine the degree to which such datasets can be used for reliable dialectometry. For this purpose, we focus on the Wancho language (Glottocode wanc1238, Hammarström et al., 2025), a Tibeto-Burman language spoken in Northeast India. Specifically, we look at the Wancho dictionary online at LivingDictionaries.app, a platform maintained by the Living Tongues institute. We focus on Wancho for a few reasons.

First, the online Wancho dictionary was created and maintained by the community themselves, despite being now largely abandoned. While linguists and anthropologists have been involved in work on Wancho for some time, the dictionary is the work of speakers working directly on their language rather than outside linguists. The dictionary is entirely community facing; while the definitions of words may be in English or Hindi, example sentences in Wancho are also often given, and headwords are provided only in the Wancho script. The script, known locally as *wan$_1$ʧo$_1$ ŋau$_2$təm$_1$* 𞋀𞋁𞋂𞋃 𞋄𞋅𞋆𞋇, is a phonemic script developed by Banwang Losu starting in 2001 (Losu 2013; 2021). It accounts for the full range of phonemic contrasts, including those not present in the dialect of the creator (Losu and Morey, 2023). However, without being able to read the script, one would not be able to use the dictionary.

Second and more importantly, the dictionary incorporates dialectal forms for the given concepts with clear phonemic transcription.[1] While Wancho is broadly divided into Upper Wancho and Lower Wancho, with an occasional mention of Middle Wancho, there are a number of small divisions within the language, including village dialects and one attested clan lect for speaker from the Khã clan. The Wancho *LivingDictionaries* project is one of only a few such dictionaries which embraces dialectal forms rather than attempting to establish or follow a standard. It is the only such case for Northeast India known to the authors.

The third reason for investigating the language through this resource is that the internal makeup of the Wancho dialects is not entirely straightforward. Wancho is an important language for understanding the internal diversity of the Patkaian languages. It is the only Patkaian language other than Chang,

[1] The Wancho alphabet specifically takes into account phonemic distinctions including those found only in certain dialects. As a result, the script can be perfectly converted to an IPA-based phonemic transcription, as was done for this study.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 116–123
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

another Southwestern variety, which preserves /k/-initial pronouns which have been lost elsewhere in the group (Jacques, 2007). Thus a better understanding of its makeup would be of benefit more broadly. It is also more readily accessible than some other closely related languages due to willingness of the community to engage in work such as pedagogical discussions around dictionary creation (van Dam, 2023).

Finally, Wancho is investigated here due the author's considerable experience with the language and the Patkaian language branch within Tibeto-Burman to which it belongs. Such experience is necessary for assessing the outcomes of computationally based analyses. Without such, certain shortcomings in the method may go unnoticed if applied to a language with which one were not already somewhat familiar. This paper presents the first detailed tree model for the Wancho language.

1.1. Language Background

Wancho is a Tibeto-Burman language belonging to the Southwestern branch of Patkaian (van Dam, 2026). Little has been written on the language. Das (1988) provides a brief wordlist and scattered lexical items throughout a sociological discussion. Brief linguistic discussions were produced, including some grammatical notes (Das Gupta, 1979; Burling and Wangsu, 1998), but little if any dialectal information is clear from either on it's own.

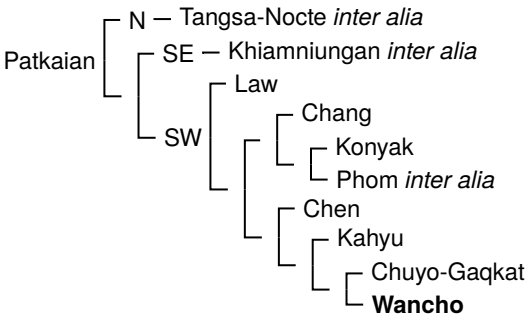


Figure 1: Place of Wancho in Patkaian

Wancho is a tonal language, with four tones based on phonation distinctions in proto Patkaian (van Dam, 2025). The typological profile is similar to closely related languages such as Konyak and Khiamniungan. There is no grammatical gender or number, although both can be explicitly marked by additional morphemes.

It is spoken almost entirely in Longding District, Arunachal Pradesh. The closest related languages are Chuyo and Gaqkat, spoken across the border in Myanmar. Some Wanchos consider these to be dialects of Wancho, and based on lexical and phonological similarities, they would be more reasonably considered as a branch of Upper Wancho. Figure 2 shows the geographic locations of the three dialect groups as traditionally given.

There is considerable dialectal diversity within Wancho, both in terms of phonology and lexicon. There are both geolects and clanlects, as well as a conventional distinction between what is called Upper Wancho and Lower Wancho. An additional Middle Wancho label is occasionally used by speakers.

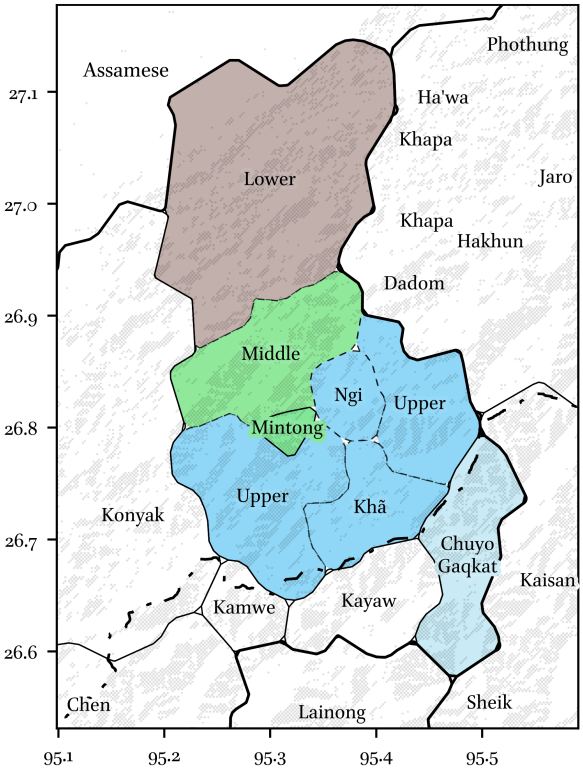


Figure 2: Locations of traditional macro groups along the Burmese border

An additional variety which falls under Wancho in terms of tribal identity is that of Mintong. It is considerably different, including its tone system and basic vocabulary such as *au* for 'mother' rather than the usual Patakaian *ɲu*, and traditional narratives hold that the residents of Mintong village settled before the arrival of the Wanchos (Losu and Morey, 2023).

Outside of the dictionary used for this study, there are two published sources of reliable lexical data the language, capturing two distinct dialects, being the wordlist of Burling & Wangsu (1998) and a follow up by Losu & Morey (2023).

2. Source data

The primary source data used for this study is scraped directly from the *LivingDictionaries* website. While the dictionary contains words from multiple dialects, many are poorly represented, with

some having only a single word. Figure 3 shows the share of each dialect, where those varieties with under 25 entries have been combined under "Others".

The majority of words in the dictionary are listed under simply "Upper Wancho" or "Lower Wancho", macro groups under which a large number of village dialects will fall. A smaller "Middle Wancho" group is also listed, although with far fewer forms given. In many cases, a single entry has been added for both a village dialect and whichever of these three macro groups under which the dialect would fall, creating multiple forms under some of the macro group labels where villages that are traditionally considered as falling under a single label differ in pronunciation. Importantly, the presence or absence of a form for a given dialect is purely of a reflection of a speaker from that dialect group adding that word.

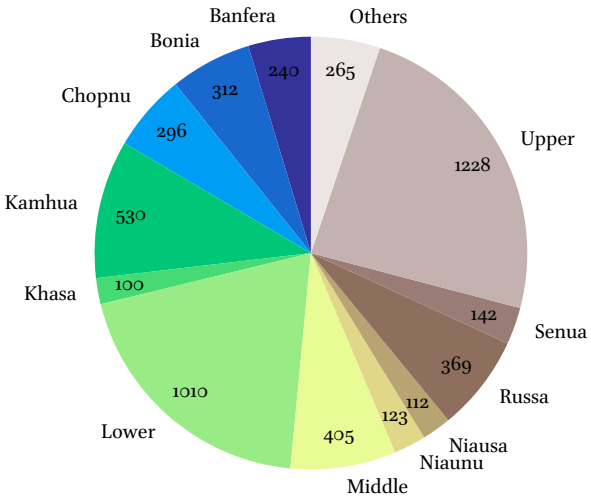


Figure 3: Dialect coverage by number of entries

We gather this data with a Python (2024) script using the *Selenium*[2] package due to the asynchronous nature of the site. The resulting data are saved as a tab-separated values (TSV) text file for further processing

2.1. Data processing

After scraping the initial dataset, some preprocessing is required prior to any analysis due to errors and inconsistencies in the data. These include example sentences in place of the English definitions, Hindi definitions in the English column, and forms which have no English entry at all but which are still recoverable based on our familiarity with the language. As an example, *𑖃𑖨𑖃 apa* is given under the kinship category for Middle Wancho, but with no definition. We know this to be the word for 'father', and so have corrected the entry in our processing. All such corrections were based on prior familiarity with the language, and then confirmed with a native speaker prior to any further analysis.

Another issue is around the inconsistency of language names. To resolve this, some doculect names are merged. For example, we combine "Hasse Russa" and "Russa" under one label, as these are either references to the same village, or references to two villages which are directly next to each other.[3] This provides a slightly better coverage for the varieties. We only do this in cases where it is certain that we are looking at one single geolect. There are also 252 entries for which no language variety is mentioned. We omit those from our analysis entirely.

For the remainder of the entries, typos in language names were corrected and forms were normalised. In many cases, multiple dialect names will be given in a single row, and are therefore flattened by separating these out so that reach row has only one dialect. We also flatten based on multiple forms given for a single word, although this is less common.

Headwords are then converted to IPA through the script, which can be done easily due to the phonemic nature of the orthography. We do not retain tones in the conversion, as they do not contribute to an analysis based on sound classes and are anyway cognate across varieties.[4] We then use the LingPy *ipa2tokens* function (List et al., 2018) to segment the IPA forms into space-separated tokens which can later be used for automatic detection of cognates using sound class alignment (SCA, List, 2012) as well as in tools such as Edictor (List and van Dam, 2024). This provides us with the final dataset for analysis.

A possible avenue for future research is to investigate the effectiveness of SCA distance on data which are only consisting of cognates, work which is currently being undertaken by the author for another language group. We do not take that step here, and instead only consider raw distances between forms provided for concepts as given. While results might be improved by only comparing across cognates, the purpose of this study is to assess the usefulness of the data with minimal

[2] github.com/SeleniumHQ/Selenium

[3] Three villages with Russa are given, being Hasse Russa, Kamkuh Russa, and Russa, belonging to the same dialect group. Kamkuh Russa and Russa are only a short walk apart, while Hasse Russa is just downstream from these two.

[4] While popular tools for computer-assisted language comparison such as LingPy (List et al., 2018) can handle tonal notation following the system of Chao (1930), we don't have contour data for each dialect. Either way, they would not add anything to the analysis undertaken here.

intervention. Using LingPy's LexStat cognate detection would not work in this case, as it relies on a large enough dataset to be able to determine correspondences versus coincidences. In the data analysed here, we have many concepts where only a handful of forms are given. It therefore lacks sufficient density of data. We actually can be certain of cognacy, however, such as that reflected in Table 1, due to familiarity with the language and having previously reconstructed various mesolects within proto Patkaian, to include proto Wancho (in preparation). However, this is based on knowledge and familiarity with the language and closely related languages which is not captured in the LivingDictionaries dataset.

DOCULECT	SCRIPT	FORM	TOKENS
Kamhua	𐑒𐑦𐑤𐑓	*ban maŋ*	b a n + m a ŋ
Upper	𐑒𐑦𐑤𐑓	*ban maŋ*	b a n + m a ŋ
Khasa	𐑏𐑦𐑤𐑓	*gan maŋ*	g a n + m a ŋ
Upper	𐑏𐑦𐑤𐑓	*gan maŋ*	g a n + m a ŋ
Chopnu	𐑝𐑦𐑤𐑓	*vən maŋ*	v ə n + m a ŋ
Banfera	𐑢𐑦𐑤𐑓	*wan mən*	w a n + m ə ŋ
Lower	𐑢𐑦𐑤𐑓	*wan mən*	w a n + m ə ŋ

Table 1: Sound class alignment for 'dream'

An alternative method would be to use SCA distance, as used by Edictor for cognate detection, but as can be seen in Tables 1, regular sound changes in Wancho and Patkaian more broadly frequently result in major changes of sound class. Thus, SCA alone does not produce accurate cognate judgements. We therefore forego analysing the data based on cognates, as SCA distance alone will still capture much of the signal that cognate identification would otherwise provide for such a small dataset.

2.2. Analysis

We model inter-dialect similarity using sound-class–based phonetic alignment distances computed with the SCA model (List, 2012) implemented in LingPy (List, 2012; List et al., 2018). IPA forms were tokenised into phonological segments and aligned using feature-informed scoring.[5] We used SCA distance rather than raw character edit distance (normalised Levenshtein distance) to better capture cognacy between forms. SCA distance is calculated similar to Levenshtein distance, but with articulatory features taken into consideration.

[5]This is done in order to ensure we are comparing similar positions within the words, and as a typical step in the workflow for using Edictor. This step is not strictly necessary, but done in order to ensure the data are more readily available for additional analyses beyond the scope of the current study.

The SCA distance is calculated by reducing sounds (IPA glyphs representing segments) into broad phonetic categories called sound classes. Velar consonants form one class, as do labials, as to Alveolar consonants, and so on.

	raŋ	*zaŋ*	*gaŋ*
raŋ	0.000	0.286	0.429
zaŋ	0.286	0.000	0.429
gaŋ	0.429	0.429	0.000

The algorithm then compares these sequences by looking up the relative cost of changing one sound class into another based on a predefined scoring matrix in the form of a directed weight graph (List, 2010, p. 39). In this matrix phonetically similar classes are inexpensive to swap and dissimilar ones are expensive. Adding or removing a sound has a greater cost than simply changing a sound to a related one. This score is then normalised into a distance between 0 and 1, where 0 represents a perfect phonetic match and 1 represents no shared phonetic traits at all. We do not make any adjustments to the weights already defined in LingPy. Table 2 shows the sound classes for the stem 'sky' *#ʒaŋ* a common Patkaian etymon.[6]

In this case, *raŋ* and *zaŋ* have a shorter distance than either do to *gaŋ*, as shown in the SCA distance matrix in Table 2.2.

DOCULECT	SCRIPT	FORM	TOKENS
Khasa	𐑏𐑓	*gaŋ*	g a ŋ
Upper	𐑏𐑓	*gaŋ*	g a ŋ
Kamhua	𐑏𐑩	*gəŋ*	g ə ŋ
Upper	𐑏𐑩	*gəŋ*	g ə ŋ
Chopnu	𐑮𐑩	*zaŋ*	z a ŋ
Senua	𐑮𐑩	*zəŋ*	z ə ŋ
Lower	𐑮𐑩̇	*zəŋ*	z ə ŋ
Lower	𐑶𐑩	*raŋ*	r a ŋ
Russa	𐑶𐑩	*raŋ*	r a ŋ

Table 2: Sound class alignment for 'sky'

Variation between /r/ and /z/ onsets is a dialectal feature by village, but coded for Lower Wancho in both ways, as the Lower Wancho forms were coded by the same contributors who added the village-specific forms. As Hasse Russa is considered a Lower Wancho dialect speaking area, the same form was added for both at the same time. This is also an important reason why forms for Upper, Middle or Lower Wancho can not be assumed to occur as such in unlisted dialects, as the macro

[6]Scriptural variations in Table 2 are due to the way in which velar codas carry the nucleus on the coda consonant. The letter 𐑓 with a dot is *-əŋ* while 𐑓 without is *-aŋ*, but then also 𐑮 is /ə/ and 𐑮 is /a/. The inclusion of the vowel is thus redundant in those forms which have it.

group forms are often just dialectal forms being over-applied.

Using the SCA analysis, pairwise dialect distances were calculated as the mean normalised alignment distance across shared lexical items. We generate a square matrix representing the linguistic distance between every pair of dialects. For every pair, it identifies the set of shared lexemes (synonyms) and calculates the average SCA distance across those forms. The resulting distance matrix was analysed using hierarchical clustering based on the unweighted pair group method with arithmetic mean (UPGMA, Sokal and Michener, 1958) and NeighborNet split network inference (Bryant and Moulton, 2004) to visualise both hierarchical structure and conflicting similarity signals. This approach situates the study within the tradition of computational dialectometry (Nerbonne and Heeringa, 1997; Heeringa, 2004; Wieling et al., 2014) while incorporating phonologically structured distance modelling.

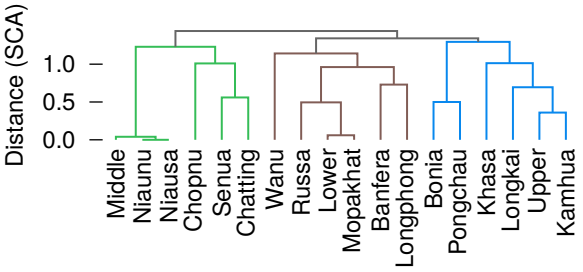


Figure 4: Neighbour joining tree with groups marked for ■ Upper, ■ Middle, and ■ Lower

The results for all dialects including the macro groups are shown in Figure 4. The tree represents the hierarchical clustering of the dialects with the inclusion of Upper, Middle and Lower. If the macro groups are removed, the overall signal of the groups is largely lost, as seen in Figure 5. This is to be expected, as in many cases, dialects such as Niau are added with Middle Wancho attached, but often as words which are not found elsewhere. Therefore, by removing the bridge between disparate concepts, we lose any sort of clear grouping which would anyway be discernible were one to simply look at the data and know the sound correspondences.

The macro groups simply have much greater representation in the source data. Additionally, different contributors have transcribed forms differently, but by adding both their version of the village's form along with the same form under one of the macro groups. Even though speakers may also add two different forms to the macro groups, such as in Table 2, the effects of such conflicting forms do not unnecessarily affect the overall connections between the macro-groups. This is discussed further below in Section 3 on data issues. We mention the results of the culled dataset here, but do not believe they are a realistic reflection of the languages' relationships given the incredibly sparse data we are left with in the absence of macro dialect labels.

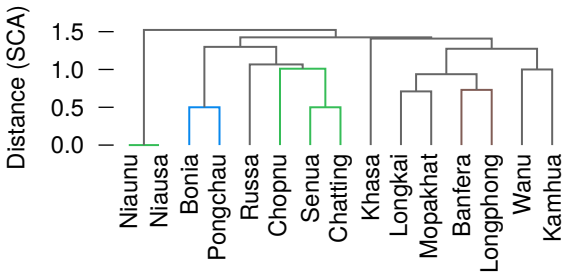


Figure 5: Neighbour joining tree without macro dialects

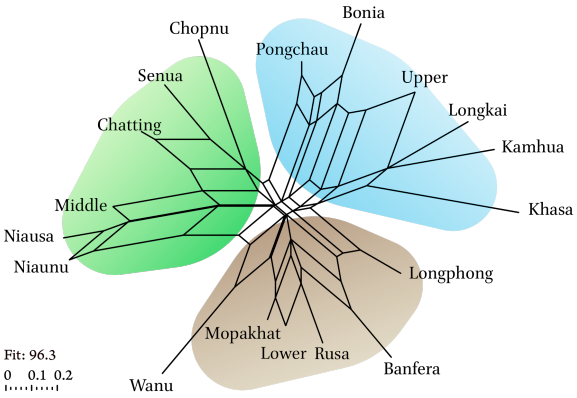


Figure 6: Neighbour net with traditional groupings of ■ Upper, ■ Middle, and ■ Lower Wancho

In general, we see the traditional three subgroups showing up well in the data, as illustrated in the neighbour joining network shown in Figure 6, in which the distances between the dialectal forms based on sound class alignments group more or less cleanly into three groups.

3. Discussion

A number of features of the data contribute to the results. First, as mentioned, many village-specific dialects are under-represented. Even after merging paired villages such as Konnu and Konsa or Niaunu and Niausa,[7] it is still not close to an even distribution.

However, this is not always a problem due to using SCA distance measures. For example, in

[7] *nu* meaning 'large' derived from 'mother' and *sa* meaning 'small' derived from 'child', a common source of augmentative and diminutive marking in the Patkaian language group to which Wancho belongs.

spite of the low number of entries for Wanu (15 items total), it is still correctly placed within Lower Wancho in the trees. This would not be the case with other approaches such as cognate coding as characters in a Bayesian phylogeny. We also see Chopnu in a stable position, as one of the more transitional varieties between Lower and Middle Wancho. Thus, we find strong signal for the Upper / Lower division frequently reported by the community, as well as a clear explanation for why Middle often comes up only occasionally as a label as a less pronounces subbranch within Lower. In cases where only one one or a few words are given, such as with Otongkhua and Lawnu, they will naturally appear as more similar to dialects for which those words are given assuming some level of shared cognacy.

For example, in Lawnu, 'give' is *koʔ*, which is also the only word for the dialect present in the dataset for that dialect. Therefore Launu will always group only with other languages which also have 'give' as *koʔ*. This will be the case regardless of where Lawnu would actually be rightfully positioned in the tree. This is especially problematic since we know that in fact *all* dialects mentioned in the LivingDictionaries dataset will have this same stem for 'give', as is the case for most of Patkaian beyond Wancho. Cognates found in related non-Patkaian groups as well, such as Jinghpaw *tʃoʔ* (Yuè, 2006). However, within the Wancho data, only a handful of dialects have been given the form for 'give', resulting in the inclusion of this one word having a much stronger influence than it ought to in terms of determining the tree model. To resolve this, we can filter the data which is being studied to only those concepts which are more widely represented in the data, such as occurring in a minimum number of dialects. A better solution is simply to remove those dialects which only have a handful of concepts, which is what we have done here. As such, only varieties with more than 5 concepts encoded in the dictionary are considered for our analysis.

On the other hand, were we to stick to only those dialects which have considerably better representation, such as those with at least 200 concepts, we would be unlikely to get out of any other views of Upper-Middle-Lower categories. That is, the results would only show that three such groups exist, with Kamhua being strictly within Upper, as nearly all Kamhua words have an identical Upper Wancho counterpart in the dataset. Such a high cutoff does not help in clarifying the internal relationships of the dialects. Instead, the obvious long-term solution is simply to incorporate more data. However, at the time of writing, there are no active contributions to the dictionary due to limits on time of some community members, and issues around access for others.[8]

Translations or other concept issues are also problematic, with many coming directly from Google searches, and at a time when Google was plagued by AI trained on Reddit (Olson and Kerr, 2024; Tong, 2024). For example, the Chopnu dialect form for 'generation' 𞋀𞋁𞋂𞋃 instead gives a summary of a 2019 computer game called Generation Zero (2019). 'Banana' 𞋄𞋅 for Senua has as a semantic domain "Units of measure", unintentionally referencing a popular meme from the online forum *Reddit*. This is one issue with a dictionary where English is the target language for definitions but may not be widely understood within the community, and for which the contributors may rely on potentially misguided Google AI search results when those were particularly precarious.

DOCULECT	SCRIPT	FORM	TOKENS		
Nginu	𞋆𞋁𞋇	*bək*	b	ə	k
Konnu	𞋆𞋁𞋇	*bək*	b	ə	k
Konsa	𞋆𞋁𞋇	*bək*	b	ə	k
Upper	𞋆𞋁𞋇	*bək*	b	ə	k
Kamhua	𞋆𞋁𞋇	*bək*	b	ə	k
Khasa	𞋅𞋁𞋇	*gək*	g	ə	k
Upper	𞋅𞋁𞋇	*gək*	g	ə	k
Kamhua	𞋉𞋁𞋇	*vək*	v	ə	k
Upper	𞋉𞋁𞋇	*vək*	v	ə	k
Pongchau	𞋌𞋁𞋇	*wək*	w	ə	k
Upper	𞋌𞋁𞋇	*wək*	w	ə	k
Russa	𞋌𞋁𞋇	*wək*	w	ə	k
Lower	𞋌𞋁𞋇	*wək*	w	ə	k

Table 3: Sound class alignment for 'pig'

Finally, there are considerable variations in how terms are transcribed, as previously mentioned above. In some cases these are clearly due to dialectal differences. In other cases it could either be the result of intra- or inter-speaker variation, an unfamiliarity with the script, or uncertainty around the phonetic realisation of a term.

As an example, the term 'pig' is given in Table 3. For Kamhua, the term is given as both 𞋆𞋁𞋇 *bək* and 𞋉𞋁𞋇 *vək*, which is a reasonable case of allophony or the transcription of a sound which may not be entirely clear to the speakers themselves. However, the same term occurs for the Upper Wancho macrogroup as both 𞋌𞋁𞋇 *wək* and 𞋅𞋁𞋇 *gək*. The former is again understandable as

[8] It should be mentioned that, at the time of writing, a new version of the dictionary is being developed for the community with less access restrictions, and in way in which such automated analyses as the one described here would be more accessible.

allophony/allography from *vək*, whereas the /g/ onset is actually a dialectal difference, found in the Khasa dialect. However, as Kamhua and Khasa are both considered Upper Wancho dialects, the "Upper Wancho" entries reflect both pronunciations as speakers from both villages added their forms as Upper Wancho forms. A more comprehensive study would need to address these issues to provide a clearer picture of the internal relationships between dialects.

Still, it is our hope that contributions resume once more. For this reason, we do not provide a full CLDF dataset at this time.

SCA will fail on clear cognates.

4. Conclusion

The analysis presented here shows that even with a fairly sparse dataset across multiple dialects, by analysing lexical data based on sound class alignments (SCA), reliable genealogies can be produced. By limiting our analysis to only those concepts which exist in at least three doculects, we prevent strictly paired entires such as Upper Wancho and Kamhua where both were added at once for the same concepts and with identical segmental forms. By only investigating concepts with at least 5 shared concepts, we avoid unnecessary noise caused by under-represented varieties.

Importantly, however, this analysis relies on a considerable amount of work to ensure that concepts are properly linked in the code. Ensuring that forms given properly match the concepts found throughout the dialects given requires some understanding of the language, both in terms of historical phonology and lexical formation strategies.

However, armed with this knowledge, we are able to produce an otherwise automated analysis that takes community-managed data as its input, analyses a community-developed phonemic script, and as a gain quantifiable evidence for internal subgrouping. This same conclusion could be gained through a comparative study with a minimal concept list were one to travel through the district and then investigating the forms provided. However, our approach shows that even in lieu of direct fieldwork or consistent methodologies around data collection, non-academic sources of data still provide an incredibly important asset for language typology.

Finally, based on the current analysis combined with discussions with native speakers and previous analyses on the language for the purposes of cross-linguistic comparisons, we propose the first detailed tree model for Wancho in Figure 7 a tree model for Wanchoic. We include Chuyo and Gaqkat, based on unpublished fieldwork data no present in the LivingDictionaries project, which show a high degree of similarity with Upper Wancho varieties. A number of lexical differences are found, moreso in Gaqkat than Chuyo, which could either indicate strata effects caused by non-Wancho Patkaian languages spoken in Myanmar, or, that Gaqkat represents an earlier form of Southwestern which has largely shifted more toward Wancho as a result of close contact with Chuyo, which is more clearly aligned with Upper Wancho.

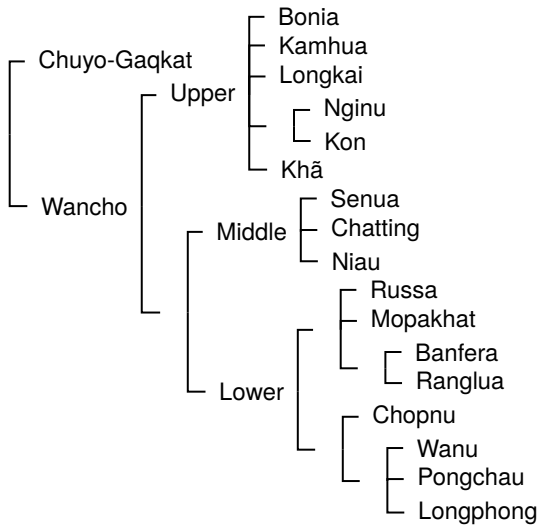


Figure 7: Internal structure of Wanchoic

This analysis is based on both the findings of the current study, along with community descriptions of language variety relatedness. Our results here show that conventional descriptions by speakers are largely accurate for Wancho, supported by both regular sound changes indicative of subbranches, as well as lexical variation for which Lower Wancho shows more similarity to neighbouring Northern Patkaian languages such as Nocte.

5. Acknowledgements

Thank you to Banwang Losu for his help confirming corrections made to the scraped data, and to the members of the Wancho Literary Mission for their work on the language. Thank you to Mattis List and Jessica Nieder for their comments on an earlier version of this paper. Finally, thank you to the anonymous reviewers for their comments and suggestions.

6. Supplemental Materials

All code and data used in this study are available on Codeberg repository at patkaiist/wancho. The original LivingDictionaries project is available at livingdictionaries.app/wancho.

7. Bibliographical References

David Bryant and Vincent Moulton. 2004. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution*, 21(2):255–265.

Robbins Burling and Mankai Wangsu. 1998. Wancho phonology and word list. *Linguistics of the Tibeto-Burman Area*, 21(2):43–71.

Yuen Ren Chao. 1930. ə Sistim əv Toun-letəz. *Le Maître Phonétique*, 30(30):24–27.

Pranjana Das. 1988. *Wancho through history*. Ph.D. thesis, North-Eastern University, Shillong.

Kamalesh Das Gupta. 1979. A note on the Wancho language of Arunachal Pradesh. *Resarun: journal of the Research Department, Government of Arunachal Pradesh*, 5(1):25–37.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, University of Groningen.

Guillaume Jacques. 2007. A shared suppletive pattern in the pronominal systems of Chang Naga and Southern Qiang. *Cahiers de linguistique Asie orientale*, 36(1):61–78.

Johann-Mattis List. 2010. SCA: Phonetic alignment based on sound classes. In *European Summer School in Logic, Language and Information*, pages 32–51. Springer.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2018. LingPy: A Python Library for Historical Linguistics. *Journal of Open Source Software*, 3(28):1242.

Johann-Mattis List and Kellen Parker van Dam. 2024. Computer-Assisted Language Comparison with EDICTOR 3. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 1–11.

Banwang Losu. 2013. *The Wancho Script*. Partridge India, New Delhi.

Banwang Losu. 2021. Phonology of the Wancho Language and Script. Master's thesis, Department of Linguistics, Deccan College, Pune. MA dissertation.

Banwang Losu and Stephen Morey. 2023. The Wancho language of Kamhua Noknu village. *Linguistics of the Tibeto-Burman Area*, 46(2):201–234.

John Nerbonne and Wilbert Heeringa. 1997. Measuring Dialect Distance Phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, pages 11–18.

Paayal Olson and Dara Kerr. 2024. Google strikes $60 million deal with Reddit, allowing search giant to train AI models on human posts. *CBS News*.

Python Software Foundation. 2024. Python 3.13. Computer software.

Robert R. Sokal and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Systemic Reaction / Avalanche Studios. 2019. Generation Zero. generationzero.com. Computer software.

Anna Tong. 2024. Why Google's AI Overviews Gets Things Wrong. *MIT Technology Review*.

Kellen Parker van Dam. 2023. Developing & Maintaining Community-Driven Language Resources: Dictionaries & Narratives. Talk presented at the Workshop on Pedagogic Development. Organised by the Wancho Literary Mission with support from the Assam Rifles, Longding, Arunachal Pradesh.

Kellen Parker van Dam. 2025. On the independence of tonogenesis in Patkaian branches. *Linguistics of the Tibeto-Burman Area*, 48(2):163–192.

Kellen Parker van Dam. 2026. Patkaian (Northern Naga). In Kristine Hildebrandt, Yankee Modi, David Peterson, and Hiroyuki Suzuki, editors, *The Oxford Guide to the Tibeto-Burman Languages*. Oxford University Press, Oxford. Forthcoming.

Martijn Wieling, John Nerbonne, and R. Harald Baayen. 2014. Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE*, 9(9):e105342.

Yuè Má Là 岳麻腊. 2006. 景颇语杜连话概况 [Overview of the Duliang Dialect of the Jingpo Language]. 民族语文, (4):68–80.

Dialectometry and Evaluation of the ePark Corpus for Low-Resource Formosan Language Dialects

Henry Gagnier
Pittsford Sutherland High School
Pittsford, NY, USA
henrygagnier9@gmail.com

Abstract

Formosan languages are a critically endangered branch of the Austronesian family spoken in Taiwan, and many of their dialects remain poorly understood and computationally understudied. Subgrouping relationships in these languages are often contested and unresolved. We provide the first evaluation of the ePark corpus as a dialectal NLP resource, identifying its strengths and gaps for future NLP work, and present the first large-scale corpus-based computational analysis of dialect similarity across all officially recognized Formosan languages. We use the ePark corpus to analyze 42 dialects in 16 Formosan languages, and through word-level TF-IDF cosine similarity, Jaccard similarity over shared vocabulary, and Levenshtein distance, we quantify pairwise dialectal relationships within the Amis, Atayal, Seediq, Bunun, Paiwan, Rukai, and Puyuma languages. We find that simple lexical similarity methods can recover and confirm linguistically established dialectal subgroupings. We find that in multiple cases the two metrics diverge, offering insights on contested subgroupings such as Mantauran Rukai. This work establishes a scalable methodological framework for dialectometry in low-resource languages, demonstrates the value of the ePark corpus for Formosan NLP research, and encourages future work in NLP on Formosan dialects.

Keywords: dialectometry, Formosan languages, dialects, ePark corpus, corpus linguistics

1. Introduction

Formosan languages are a primary branch of the Austronesian language family spoken by under 200,000 people in Taiwan (Lin et al., 2025; Li, 2008). Despite their importance to historical linguistics (Zeitoun and Goudin, 2024), most Formosan languages are severely under-resourced and endangered, with many exhibiting substantial internal variation. Formosan dialects often differ in lexicon, phonology, and morphosyntax, reflecting complex histories of migration, contact, and isolation (Li, 2014). Understanding relationships between these dialects is crucial to linguistic classification and documentation and supports the development of language technology that can enable preservation, revitalization, and education efforts for Formosan languages (Zeitoun et al., 2003; Meighan, 2021).

Computational methods in dialectology known as dialectometry have emerged as a promising method for dialect analysis in low-resource languages, aggregating a large number of individual linguistic features to produce satisfactory results (Bompolas, 2023). Rama et al. (2017) uses Levenshtein distance and autoencoders to analyze Gondi dialects. (Dunn, 2021) analyzes the dialects of high-resource languages using two Spearman frequency-based similarity measures across multiple domains. (Kwaik et al., 2018) uses the Jaccard similarity, a vector space model (VSM), latent semantic indexing, Hellinger distance, and correlation coefficients between frequent words. Bompolas and Melissaropoulou (2025) uses grammatical and lexical features to quantify dialect differences in Inner Asia Minor Greek varieties. These computational methods provide extremely important insights, especially for dialects without current support in natural language processing (NLP).

Research on Formosan languages in NLP is beginning but is still extremely limited. Zheng et al. (2024) investigates machine translation (MT) methods for low-resource Formosan languages. (Lin et al., 2025) develops FormosanBench, a benchmark for MT, automatic speech recognition (ASR), and summarization in Atayal, Paiwan, and Amis, and finds that existing LLMs perform poorly on Formosan NLP tasks. (Hsieh et al., 2024) uses self-supervised models to obtain and analyze their speech representations and develop a model to classify Formosan languages, and uses the speech representations to construct a linguistic phylogeny of Formosan languages. While work has begun in dialectometry and Formosan languages, Formosan dialects and their dialectal relationships have not been studied in NLP. Considering the importance of dialectometric analysis and the absence of work on Formosan language dialects in NLP, it is necessary to begin computational work on these under-resourced Formosan dialects.

The purpose of this study is to (1) provide the first large-scale corpus-based computational measurement of dialect similarity across all major For-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 124–134
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

mosan languages, (2) test whether lexical similarity measures recover relationships established in prior research, and (3) provide an evaluation of the ePark corpus as a dialectal resource. This study also aims to improve the inclusion of low-resource Formosan languages and dialects in NLP research and contribute to the evaluation of dialectal corpus resources, particularly the ePark corpus, as a NLP resource for language and dialect analysis.

2. Methodology

2.1. Data

We use the ePark corpus[1] (Mohamed et al., 2024; Indigenous Languages Research and Development Foundation, 2020), a high-quality, comprehensive resource for the preservation, learning, and revitalization of Indigenous languages of Taiwan. It was developed by the Indigenous Languages Research and Development Foundation and covers 42 official dialects of the 16 different Formosan languages being Amis, Atayal, Saisiyat, Thao, Seediq, Bunun, Paiwan, Rukai, Truku, Kavalan, Tsou, Kanakanavu, Saaroa, Puyuma, Yami, and Sakizaya. The corpus is split into sections for vocabulary, situational dialogue, cultural texts, short passages, daily conversation, and reading and writing, and is composed of raw Formosan language text.

The full corpus contains 2680175 tokens and 587 hours of audio and translations into Mandarin, enabling language preservation, education, and research. For each dialect, we take all the text from the six sections and use this text for analysis.

2.2. Preprocessing

For each dialect corpus, we convert all text to lowercase, remove punctuation by replacing non-alphanumeric characters with whitespace with the exception of and ’ as these are used in multiple Formosan languages. We then normalize all whitespace by collapsing runs of whitespace to a single space and removing leading and trailing whitespace. Finally, we perform tokenization by splitting text on boundaries for word-level tokens. Reliable morphological analysers do not exist for most Formosan languages, so we do not perform any segmentation beyond word-level tokenisation.

2.3. Lexical Similarity Measures

We compute three lexical similarity measures for each pair of dialect varieties within a language group, excluding identical pairs.

[1] https://ai4commsci.gitbook.io/formosanbank/

First, we compute word-level TF-IDF cosine similarity to capture the weighted similarity of each dialect. We implemented TF-IDF vectors and cosine similarity using `scikit-learn` (Pedregosa et al., 2011). We set `min_df` to 2 in order to exclude terms appearing in only one document and limit the effect of noise that likely does not affect dialect-level differences. We set `sublinear_tf` to True in order to reduce the influence of high-frequency words, which tend to carry little discriminative information for dialect analysis (Manning et al., 2008).

We then compute Jaccard similarity over the presence of word types to measure unweighted vocabulary overlap. We reduce noise from low-frequency terms, which are likely to be noise or corpus-specific artifacts, by restricting the vocabulary to words with a corpus frequency of at least 5 in each dialect.

Third, we compute character-level Levenshtein distance on the 500 most frequent word types in each dialect, averaged symmetrically across both directions and normalized by word length to capture formal similarity between word types beyond the binary presence/absence distinctions made by Jaccard.

To assess the statistical reliability of our similarity estimates and control for corpus size variation across dialects, we compute bootstrap confidence intervals for each language group. For each dialect pair, we draw 500 bootstrap samples of 10,000 tokens with replacement from each dialect’s token list, fixing the sample size across all pairs to ensure comparability. We recompute TF-IDF cosine similarity, Jaccard similarity, and Levenshtein for each sample and report the mean width of the 95% confidence interval across all pairs within each language group as a measure of estimate stability. To encourage further research on low-resource Formosan dialects and increase reproducibility, we release code at https://github.com/henrygagnier/formosan-dialectometry.

3. Results

3.1. Corpus Statistics

We first look at the overall corpus statistics from ePark (Table 1), which summarize the size and diversity of each dialect. Token counts and types across all dialects and languages are comparable, ranging from 46,177 to 67,968 and 9,530 to 13,390, respectively. Token-type ratios are also fairly consistent, ranging from 0.184 to 0.240. Seven of the sixteen languages have multiple officially recognized dialects. Hapax rates are high across all dialects, ranging from 0.721 to 0.793, indicating that a large proportion of each dialect’s vo-

cabulary is represented by a single occurrence in the corpus. This sparsity is a corpus-wide property and motivates the restriction of Jaccard similarity to types with a frequency greater than 5.

The vocabulary section constitutes between 21.9% and 32.6% of tokens per dialect, suggesting that section composition does not introduce large cross-dialect imbalances due to the vocab size consistency. Among languages with multiple dialects, the shared core vocabulary of the types appearing in all dialects at a frequency greater than 5 varies from 8.8% to 24.5%. The dialect-unique vocabulary, or types that appear in only one dialect ranges from 53.8% to 79.2%. This implies that languages with higher shared core percentages such as Paiwan and Seediq are good candidates for cross-dialect model transfer, while Atayal and Rukai's high dialect-unique percentage suggest the dialect-specific resources are necessary for effective language technology.

3.2. Dialect Similarity

We now look at the overall similarity of each Formosan language with multiple dialects (Table 2). We find that the Seediq dialects are the most similar among the languages based on cosine similarity, with a mean cosine similarity of 0.820. We also find that the Atayal and Rukai dialects differ the most among Formosan languages, with a mean cosine similarity of 0.605 and 0.648, respectively. Paiwan, Amis, Puyuma, and Bunun all have moderate variance, with the mean cosine similarity ranging from 0.676 to 0.781 and the mean Jaccard similarity ranging from 0.333 to 0.447. Bunun, Puyuma, Amis, Paiwan, and Seediq all had relatively similar Levenshtein distances ranging from 0.830 to 0.862. Rukai and Atayal had lower Levenshtein distances of 0.761 and 0.792, respectively.

Looking at the cosine similarity by section (Table 3), we see that the vocabulary section is the most similar in all languages, with cosine similarity scores of 0.811 to 0.930, while the cultural and reading and writing sections are the most different, with cosine similarity scores between 0.412 and 0.672 and 0.425 and 0.708, respectively. Due to the large differences in dialectal similarity in cultural, reading and writing, and daily conversation topics, in dialectal analysis, similarity and distance scores must be compared only with scores computed using the same set of topics, and NLP work must consider these differences.

We now zoom into the per-dialect pairwise TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance for each Formosan language (Figure 1).

In Amis, it can be seen that the Central, Chengkung-Kwangshan, and Southern dialects are similar with pairwise cosine values ranging from 0.773 to 0.828. The Northern Amis dialect is an outlier, with cosine similarities to other dialects ranging from 0.640 to 0.684, and its Jaccard similarities with other dialects are lower than other pairs, ranging from 0.323 to 0.345. Tavalong-Vataan has moderate relative cosine similarity with other languages, ranging from 0.736 to 0.793. This computational finding aligns with Li and Tsuchida (2022) which finds that Amis dialects in the north are far more diverse than central and southern Amis dialects. Li and Tsuchida (2022) also finds that the Central and Southern Amis dialects are extremely close to each other and that Tavalong-Vataan can be grouped with either the Northern or Southern dialects through lexical differences. Atayal has the lowest mean cosine similarity of the Formosan languages, reflecting high diversity among its dialects. We find that the Skikun-Squliq and the Matu'uwal-Plngawan dialect pairs have the high cosine similarities of 0.728 and 0.719, respectively. Interestingly, while the Skikun-Squliq fair has a high cosine similarity and Jaccard similarity, the Matu'uwal-Plngawan pair has a low Jaccard similarity of 0.238. Other pairs with relatively high cosine similarity include S'uli-Klesan and S'uli-Skikun. The Matu'uwal-Klesan pair has the lowest cosine similarity of 0.526. (Goderich, 2020) groups Atayal into two main subgroups: Northern Atayal and Southern Atayal. Northern Atayal is further grouped into Matu'uwal and Nuclear Northern Atayal (Skikun and Squliq), and Southern Atayal is further grouped into Plngawan and Nuclear Southern Atayal (Klesan, S'uli). The high cosine similarity of Skikun-Squliq is consistent with the classification of the dialects as Nuclear Northern Atayal, which sets them apart from Matu'uwal. Despite being a member of the proposed Northern Atayal subgroup, Matu'uwal has a fairly low cosine similarity with the other members of the Northern Atayal group and a high cosine similarity with Plngawan, a Southern Atayal dialect. Conversely, the low Jaccard similarity and Levenshtein distance between Matu'uwal supports its grouping as a non-Northern Atayal dialect.

Seediq dialects have the highest mean cosine similarity of any Formosan language of 0.820, indicating that dialects are fairly similar. Toda and Truku are the most similar, with a cosine similarity of 0.895 and a high Jaccard similarity of 0.425. The Tgdaya-Truku and Tgdaya-Toda pairs have a lower cosine similarity of 0.765 and 0.799, respectively. While lower, these dialects are still fairly similar compared to dialect similarities in other Formosan languages. (Ogawa and Asai, 1935) identifies Paran/Tgdaya Seediq and Truku Seediq as the two main dialectal groups of Seediq, with Truku Seediq encompassing Truku and Toda. Our computational findings match this finding, with Toda

Language	Dialect	Tokens	Types	TTR	Hapax rate	Vocab%	Core%	Uniq%
Amis	Northern	47,885	9,979	0.208	0.791	31.6	19.6	53.8
	Tavalong-Vataan	51,790	10,767	0.208	0.774	29.7		
	Central	50,338	10,341	0.205	0.770	29.9		
	Chengkung-Kwangshan	46,967	10,315	0.220	0.793	31.9		
	Southern	50,603	10,727	0.212	0.771	30.1		
Atayal	Squliq	56,282	10,853	0.193	0.737	26.8	8.8	71.0
	S'uli	54,378	10,274	0.189	0.742	28.0		
	Matu'uwal	57,312	10,819	0.189	0.756	27.2		
	Plngawan	46,177	9,009	0.195	0.764	32.6		
	Skikun	57,914	10,824	0.187	0.728	26.1		
	Klesan	54,355	10,200	0.188	0.779	28.1		
Saisiyat	Saisiyat	52,442	9,662	0.184	0.759	29.5	—	—
Thao	Thao	49,846	9,530	0.191	0.750	29.4	—	—
Seediq	Toda	51,636	9,957	0.193	0.750	29.3	23.8	61.5
	Tgdaya	51,686	10,028	0.194	0.763	29.1		
	Truku	58,746	10,872	0.185	0.748	25.7		
Bunun	Takbunuaz	49,958	10,715	0.214	0.774	30.4	15.2	61.0
	Takivatan	51,953	11,107	0.214	0.772	29.2		
	Takituduh	47,699	10,391	0.218	0.784	32.1		
	Takibaka	51,127	10,526	0.206	0.771	30.3		
	Isbukun	47,868	10,409	0.217	0.774	31.6		
Paiwan	Southern/Central	49,376	10,536	0.213	0.770	30.4	24.5	56.7
	Northern	59,902	12,633	0.211	0.769	25.3		
	Central	59,479	12,353	0.208	0.767	25.5		
	Eastern	60,056	12,745	0.212	0.781	25.5		
Rukai	Tanan	48,165	10,600	0.220	0.778	31.3	9.3	79.2
	Budai	47,800	10,835	0.227	0.796	31.4		
	Labuan	48,737	10,634	0.218	0.769	30.8		
	Tona	52,733	11,469	0.217	0.769	29.0		
	Maga	50,614	12,168	0.240	0.781	29.8		
	Mantauran	56,976	12,620	0.221	0.777	26.6		
Truku	Truku	51,578	9,747	0.189	0.754	29.3	—	—
Kavalan	Kavalan	57,141	11,026	0.193	0.721	25.7	—	—
Tsou	Tsou	59,577	10,972	0.184	0.740	25.9	—	—
Kanakanavu	Kanakanavu	51,278	10,467	0.204	0.746	29.4	—	—
Saaroa	Saaroa	56,707	12,492	0.220	0.763	26.6	—	—
Puyuma	Nanwang	67,968	13,390	0.197	0.740	21.9	14.8	63.9
	Katipul	56,315	11,186	0.199	0.748	27.0		
	Kasavakan	56,943	11,051	0.194	0.733	26.8		
	Ulivelivek	52,773	10,839	0.205	0.760	28.9		
Yami	Yami	52,209	10,894	0.209	0.775	26.7	—	—
Sakizaya	Sakizaya	50,565	10,477	0.207	0.773	29.9	—	—

Table 1: Corpus statistics and resource evaluation metrics for all Formosan languages and dialects in the ePark corpus.

and Truku being extremely similar and Tgdaya being relatively different from the Truku Seediq dialects.

Bunun dialects have a moderate relative variance with a mean cosine similarity of 0.676. Computationally, we find that the Takivatan-Takbunuaz and Takituduh-Takibaka dialect pairs have the highest cosine similarities of 0.774 and 0.767, respectively. Moreover, these two pairs have the highest Jaccard similarity in Bunun as well. Other pairs have low or moderate relative similarity with cosine similarities ranging from 0.619 to 0.674. Ogawa and Asai (1935) proposes three subgroups of Bunun: Southern Bunun (Ishbukun), Central Bunun (Takbanuaz and Takivatan), and Northern Bunun (Takibaka, Takituduh). Our results confirm this grouping with extreme precision, with the Takivatan-Takbunuaz and Takituduh-Takibaka

		Cosine similarity			Jaccard similarity			Levenshtein distance			
Language	**N**	Mean	SD	CI	Mean	SD	CI	Mean	SD	CI	**Core%**
Seediq	3	0.820	0.055	0.018	0.362	0.047	0.015	0.848	0.024	0.005	23.8
Paiwan	4	0.781	0.015	0.019	0.434	0.015	0.014	0.861	0.011	0.004	24.5
Amis	5	0.728	0.064	0.019	0.447	0.098	0.015	0.862	0.035	0.007	19.6
Puyuma	4	0.726	0.068	0.017	0.333	0.087	0.014	0.830	0.042	0.005	14.8
Bunun	5	0.676	0.051	0.018	0.366	0.041	0.014	0.844	0.016	0.006	15.2
Rukai	6	0.648	0.063	0.014	0.265	0.048	0.012	0.761	0.038	0.006	9.3
Atayal	6	0.605	0.062	0.016	0.270	0.045	0.013	0.792	0.028	0.007	8.8

Table 2: Summary of pairwise dialect similarity by language group

Language	Vocabulary	Situational	Cultural	Short pass.	Daily conv.	Read/Write	All sections
Seediq	**0.930**	0.876	0.672	0.881	0.733	0.708	0.820
Paiwan	**0.888**	0.845	0.603	0.816	0.661	0.634	0.781
Amis	**0.853**	0.771	0.534	0.777	0.559	0.527	0.728
Puyuma	**0.861**	0.788	0.576	0.748	0.612	0.581	0.726
Bunun	**0.847**	0.724	0.520	0.727	0.484	0.447	0.676
Rukai	**0.847**	0.727	0.439	0.696	0.465	0.439	0.648
Atayal	**0.811**	0.662	0.412	0.643	0.455	0.425	0.605
Mean	*0.863*	*0.770*	*0.537*	*0.755*	*0.567*	*0.537*	*0.712*

Table 3: Mean pairwise TF-IDF cosine similarity between dialects, computed separately for each ePark corpus section and across all sections combined

pairs lining up exactly with the Central Bunun and Northern Bunun pairs. This precise recovery of an established three-subgroup structure from lexical similarity demonstrates the potential of corpus-based dialectometry as a scalable complement to traditional methods in dialectology.

Paiwan's dialects are very similar, with a mean cosine similarity of 0.781. The differences in similarity in dialect pairs are extremely small, with a standard deviation of the cosine similarity being 0.015 and cosine similarity ranging from 0.765 to 0.802. (Ho, 1978) proposed the grouping of Paiwan in Southeast and Northwest Li (1999) concluded that the internal relationship of Paiwan still remains unclear. Computationally, we are unable to make conclusions regarding the subgrouping of Paiwan, as all dialect pairs have similar cosine similarity.

Rukai dialects vary greatly, with a mean cosine similarity of 0.648 and the lowest mean Jaccard similarity of all Formosan languages of 0.265. The Budai and Labuan dialects have a high similarity of 0.802. Tona, Mantauran, and Maga have relatively moderate to high cosine similarities with each other, ranging from 0.674 to 0.714. Looking at Jaccard similarity, different trends emerge. Budai and Labuan are consistently the most similar, while Tona, Mantauran, and Maga have low Jaccard similarities. Labuan, Tanan, and Budai have moderate-high Jaccard similarities, and Mantauran is an outlier with low Jaccard similarity with all other Rukai dialects. Dialectal relationships among Rukai dialects are under debate, with Tanan, Budai, and Labuan being designated as the Eastern Rukai Group and Tona and Maga in the Western Rukai Group (Hsin, 2000; Chen, 2011). Computationally, through Jaccard similarity, we are able to see the Western Rukai Group clearly. We are also able to see the Eastern Rukai Group using cosine similarity. The position of Mantauran is controversial, with hypotheses that Mantauran groups with Tona and Maga (Li, 1997), Mantauran is the first offshoot of the Rukai family (Starosta, 1994, 1995), and that Mantauran forms a subgroup with Tanan, Budai, and Labuan (Zeitoun, 2007). We find that Mantauran aligns with the Western Rukai Group through cosine similarity. In contrast, we find that Mantauran does not align with either group using Jaccard similarity and Levenshtein distance, supporting Starosta (1994).

Puyuma has moderate dialectal variation with a mean cosine similarity of 0.726. We find that the Katipul and Kasavakan dialects are the most similar, with a cosine similarity of 0.821. The Katipul-Ulivelivek and Kasavakan-Ulivelivek pairs also have high cosine similarity and Jaccard similarity. The Puyuma Nanwang dialect stands out as an outlier with low cosine similarity scores and low Jaccard similarity scores. This matches with established work, which groups the Nanwang dialect as an early offshoot of the Puyuma and as the most conservative Puyuma dialect (kuei Li, 1991; hsin

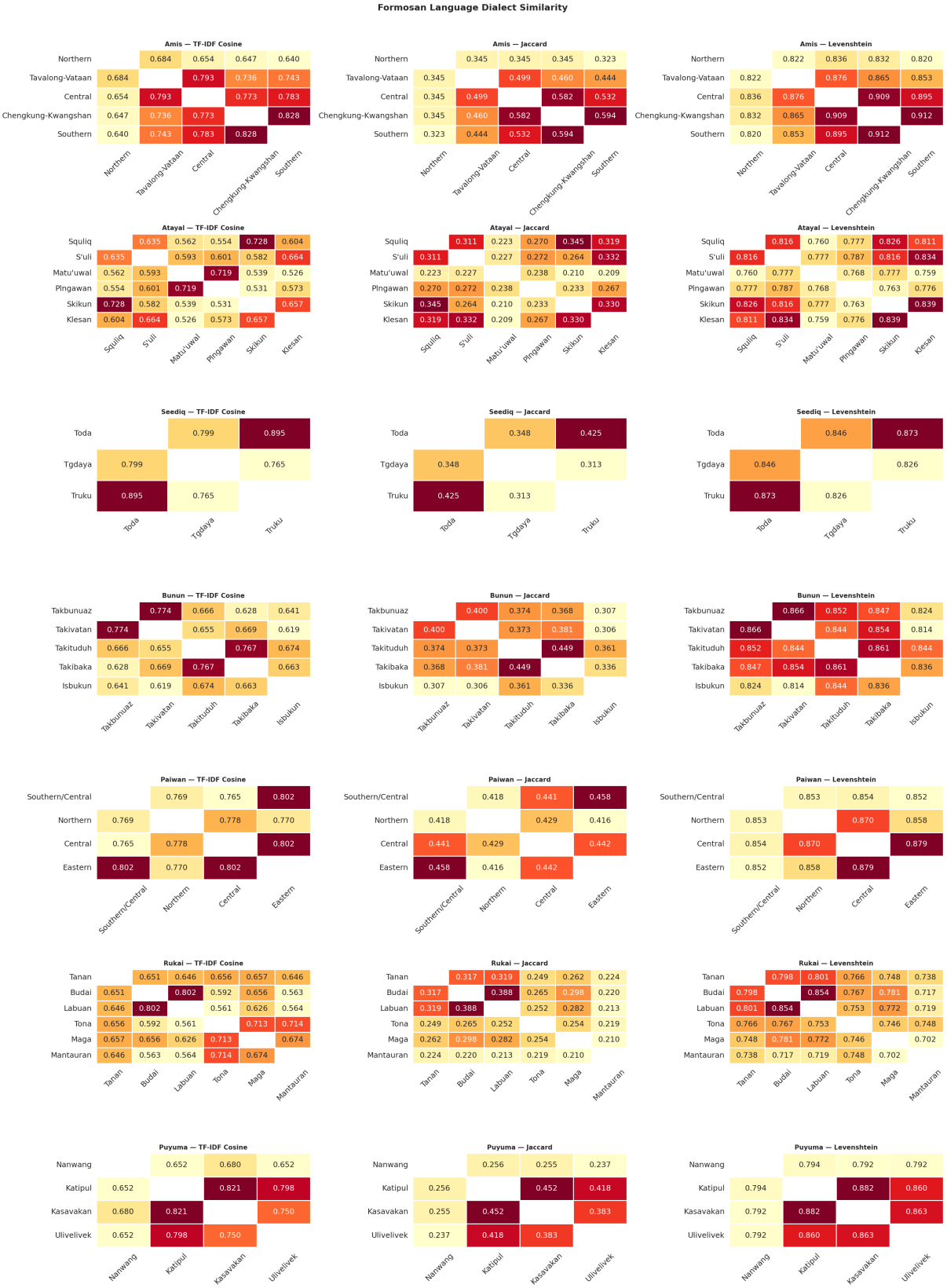


Figure 1: Per-dialect TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance for all seven Formosan languages with multiple dialects

Ting, 1978).

4. Discussion

4.1. ePark as a Dialectal NLP Resource

The ePark corpus was well-suited to this study in many ways. ePark is publicly accessible, which is

significant for low-resource dialect research when resources are often difficult to obtain. It comprehensively offers data on all Formosan languages and dialects in the same format, with text separated into six thematic sections, allowing for accurate comparisons across dialects without bias from domain differences. Each dialect had a similar token count, ensuring that imbalances in size did not greatly affect results. The corpus also focuses on language education and revitalization, meaning that it prioritizes everyday vocabulary, which is useful for dialect comparison.

However, our analysis identifies limitations that researchers should consider. The consistently high hapax rates indicate that the corpus captures a relatively shallow slice of each dialect's vocabulary. Many word types are attested only once, limiting the reliability of frequency-based similarity measures and making morphological analysis, which would require multiple attestations of inflected forms, especially difficult, unless measures are taken to mitigate these issues, such as only analyzing words of moderate and high frequency. The vocabulary section contributes such a high percentage of the tokens in each corpus, and similarity scores are highest among vocabulary, meaning that the vocabulary may overrepresent shared core vocabulary. The goals of education and language revitalization in ePark may present some challenges. Text in the corpora is curated and may underrepresent regional vocabulary that would be useful for measuring dialect similarity. The corpus also only contains official dialects and languages, meaning that all Formosan language communities and dialects may not be represented. While not utilized in this study, the ePark corpus provides audio for all text corpora, enabling further dialect research. Despite these limitations, the section structure of ePark is an advantage, allowing researchers to control for potential biases and obtain similarity estimates on text closer to naturalistic speech.

Based on our evaluation, we offer four recommendations for researchers using the ePark corpus for dialectal NLP work. First, for dialect similarity and dialect identification tasks, we recommend using the full corpus or situational dialogue and short passage sections, as analysis of the vocabulary section may underrepresent differences across Formosan dialects. Second, researchers should apply minimum frequency thresholds when computing vocabulary-based metrics due to the high hapax rates, meaning that many low-frequency types may be corpus noise rather than dialect signals. Third, for languages with a high percentage of unique vocabulary, dialect-specific models or fine-tuning are recommended instead of cross-dialect transfer, while in languages with high shared core percentages, transfer learning should be tested. Fourth, the audio data provided in the ePark corpus for all text corpora is a significantly underutilized resource that can enable phonological and ASR research in future work. We recommend ePark as a strong resource for Formosan dialectal NLP, particularly for dialect similarity modeling, dialect identification, and transfer learning experiments, while we note that tasks that require morphological annotation and speaker metadata will need supplementary resources.

4.2. Lexical Similarity as a Tool for Formosan Dialectometry

We perform the first corpus-based lexical computational analysis of dialect similarity for Formosan languages. We find that this approach can effectively capture dialect relationships in Formosan languages at scale. Using TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance computed on the ePark corpus, we find and quantify meaningful variation within and across language groups. Seediq dialects are the most internally similar, while Atayal and Rukai have the greatest internal diversity. These findings have implications for NLP development for low-resource Formosan languages and their dialects. Dialect-transfer approaches may work better for languages with high internal similarity, such as Seediq and Paiwan. Conversely, outlier dialects such as Northern Amis and highly differing languages may require dialect-specific resources for effective NLP. This work also displays the value of the ePark corpus that enables quantitative linguistic analysis on low-resource dialects and Formosan languages.

Through this study, we sought to assess whether computational lexical similarity measures recover subgrouping established through traditional linguistic methods. Among most groups, our results are extremely close to prior scholarship. For Amis, our finding that Northern Amis is an outlier from Central and Southern dialects is consistent with Li and Tsuchida (2022). For Seediq, the high similarity between Toda and Truku dialects and the divergence of Tgdaya matches the two-branch grouping proposed by (Ogawa and Asai, 1935). In Bunun, our results recover the three-subgroup structure of Southern, Central, and Northern Bunun proposed by Ogawa and Asai (1935) with precision as the highly similar Takivatan-Takbunuaz and Takituduh-Takibaka pairs correspond to Central and Northern Bunun. For Puyuma, the outlier status of Nanwang aligns with its characterization as an early offshoot and the most conservative Puyuma dialect (kuei Li,

1991; hsin Ting, 1978). These similarities give empirical support to the validity of corpus-based computational approaches as a complement to traditional methods in Formosan linguistics.

In Atayal, Matu'uwal has a high cosine similarity with Plngawan, a Southern Atayal dialect, despite its classification as a Northern Atayal dialect (Goderich, 2020). Matu'uwal's low Jaccard similarity with Plngawan suggests that while the two dialects may share high-frequency vocabulary, their overall vocabulary is limited. The divergence between cosine similarity, Jaccard similarity, and Levenshtein similarity demonstrates that each metric captures complementary aspects of lexical similarity, and that neither measure provides a full picture of dialect relationships. Cosine similarity, weighted by TF-IDF, is more sensitive to frequent shared words, while Jaccard similarity captures the broader vocabulary distribution equally. This suggests that dialects can share a common high-frequency core while diverging substantially in their wider lexicons. In Rukai, the position of Mantauran in the Rukai dialect subgrouping is contested in previous work. Our results do not resolve this debate but offer partial evidence as cosine similarity places Mantauran with the Western Rukai Group, while Jaccard similarity and Levenshtein similarity sets Mantauran apart from other Rukai dialects, which is consistent with proposals that Mantauran is the first offshoot of the Rukai language Starosta (1994). For Paiwan, dialect pairs are extremely similar, so we were unable to resolve Paiwan's structure computationally, attesting to the difficulty of determining Paiwan subgroupings (Li, 1999).

We provide the first large-scale computation analysis of dialect similarity across all recognized Formosan languages and dialects, using the ePark corpus to compute TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance. Our results confirm subgroupings established in prior linguistic research while providing computational insight into unresolved Formosan dialect subgroupings, suggesting that these methods can validate and complement traditional linguistic approaches. We hope to further encourage work on Formosan languages and dialects as an important but underresearched domain in NLP research, and we encourage researchers to use the ePark corpus for future work on low-resource language technologies.

5. Future Work

In future work, phonological and morphological features should be incorporated into computational approaches, which may allow for more accurate similarity estimates for languages where cosine similarity and Jaccard similarity are in disagreement. Formosan language morphological analysis tools should be developed, which would enable future research on low-resource Formosan dialects. These differences between cosine and Jaccard similarity in many cases should be investigated to see if they stem from linguistic properties or corpus composition. Computational analysis on Formosan dialects should include syntactic or pragmatic features to provide a more complete view of Formosan dialect relationships computationally. This work should also be extended to support downstream NLP tasks such as dialect identification, cross-dialect transfer learning, and automatic speech recognition (ASR), where the similarity estimated could inform models for extremely under-resourced Formosan language dialects. Future work should also compare corpus-based lexical similarity measures against neural and embedding-based approaches, such as multilingual language model representations, to assess whether more expressive similarity measures provide additional insight into Formosan dialect relationships. Corpora of dialect-specific text should continue to be developed and evaluated, especially for Formosan languages, due to the importance of researching and supporting dialects in NLP.

6. Conclusion

This study presents the first computational analysis of dialect similarity in low-resource Formosan languages and an evaluation of the ePark corpus for dialect analysis. We computed word-level TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance over the presence of word types to analyze dialectal relations across all recognized Formosan language dialects using data from the ePark corpus.

We find that relatively simple, corpus-based lexical similarity measures can create meaningful insights even in low-resource and endangered language settings. Our findings show that dialectometry can recover established subgroupings and further work on cases of unresolved or contested subgroupings. The divergence in the two lexical similarity methods demonstrates the importance of using multiple metrics when computationally analyzing dialect relationships. This work also establishes a scalable methodological framework for future computational dialectal analysis on Formosan and other low-resource languages, providing a step towards endangered language documentation with modern NLP techniques.

This research creates a foundation for Formosan language dialects in NLP, strengthens and complements traditional linguistics work on For-

mosan dialects. We recommend ePark as a starting point for Formosan dialectal NLP, particularly for dialect similarity modeling and dialect identification, while noting that downstream tasks requiring morphological annotation or speaker metadata will need supplementary resources. These findings and evaluations contribute to existing work on dialectometry for Formosan languages and advance the inclusion of low-resource Formosan languages and dialects in NLP research.

7. Limitations

There are several limitations that should be acknowledged in this study. First, our analysis relies on lexical similarity measures, which capture one dimension of linguistic relatedness. which may not completely reflect the underlying linguistic structure. Second, we do not perform morphological segmentation beyond word-level tokenization, which may lead to similar tokens being treated differently and limit estimate accuracy. Finally, our study covered the 16 officially recognized Formosan languages and their 42 officially recognized dialects in ePark. This official classification may not align with community-based and linguistic distinctions and varieties that are not officially recognized, and are not included in this analysis, limiting the generalizability of results in this study to all Formosan dialects.

8. Bibliographical References

Stavros Bompolas. 2023. *Computational Dialectology in the Linguistic Varieties of Cappadocian, Pharasiot, and Silliot*. Ph.D. thesis, University of Patras.

Stavros Bompolas and Dimitra Melissaropoulou. 2025. Understanding Dialectal Variation in Contact Scenarios Through Dialectometry: Insights from Inner Asia Minor Greek. *Languages*, 10(1):13.

C. M. Chen. 2011. Phonetic Evidence for the Contact-Induced Prosody in Budai Rukai. *Concentric: Studies in Linguistics*, 37(2):123–154.

Jonathan Dunn. 2021. Representations of Language Varieties Are Reliable Given Corpus Similarity Measures. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–38, Kyiv, Ukraine. Association for Computational Linguistics.

Andre Goderich. 2020. *Atayal Phonology, Reconstruction, and Subgrouping*. PhD dissertation, Institute of Linguistics, National Tsing Hua University, Hsinchu, Taiwan. Advisors: Hsiu-Chuan Liao and Hui-Chuan J. Huang.

Dah-an Ho. 1978. A Preliminary Comparative Study of Five Paiwan Dialects. *Bulletin of the Institute of History and Philology, Academia Sinica*, 49(4):565–681.

Shu-Kai Hsieh, Yu-Hsiang Tseng, Da-Chen Lian, and Chi-Wei Wang. 2024. Self-Supervised Learning for Formosan Speech Representation and Linguistic Phylogeny. *Frontiers in Language Sciences*, 3.

T. H. Hsin. 2000. *Aspects of Maga Rukai Phonology*. Ph.d. dissertation, University of Connecticut.

Pang hsin Ting. 1978. Reconstruction of Proto-Puyuma Phonology. *Bulletin of the Institute of History and Philology, Academia Sinica*, 49:321–392. In Chinese.

Paul Jen kuei Li. 1991. *Orthographic Systems for Formosan Languages*. Ministry of Education, Taipei.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A Lexical Distance Study of Arabic Dialects. *Procedia Computer Science*, 142:2–13.

Paul J. Li. 1997. *Formosa Folkways*. Chang-ming, Taipei, Taiwan. In Chinese.

Paul Jen-Kuei Li. 1999. *History of Formosan Natives: Linguistic Perspective*. Taiwan Literature Committee, Nantou, Taiwan. In Chinese.

Paul Jen-kuei Li. 2008. The Great Diversity of Formosan Languages. *Language and Linguistics*, 9(3):523–546.

Paul Jen-kuei Li. 2014. Semantic Shift and Variation in Formosan Languages. *Language and Linguistics*, 15(4):465–477.

Paul Jen-kuei Li and Shigeru Tsuchida. 2022. Subclassification of Amis Dialects. 臺灣語文研究, 17(1).

Kaiying Kevin Lin, Hsi-Yu Chen, and Haopeng Zhang. 2025. FormosanBench: Benchmarking Low-Resource Austronesian Languages in the Era of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16527–16539, Suzhou, China. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Paul J. Meighan. 2021. Decolonizing the Digital Landscape: The Role of Technology in Indigenous Language Revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17(3):397–405.

Naoyoshi Ogawa and Erin Asai. 1935. *The Myths and Traditions of the Formosan Native Tribes*. Taihoku Imperial University, Taipei. In Japanese.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Taraka Rama, Çağrı Çöltekin, and Pavel Sofroniev. 2017. Computational Analysis of Gondi Dialects. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 26–35, Valencia, Spain. Association for Computational Linguistics.

Stanley Starosta. 1994. Proto-Rukai-Tsouic: Subgroup or Treetop? In *Proceedings of the Seventh International Conference on Austronesian Linguistics*, Leiden, the Netherlands.

Stanley Starosta. 1995. A Grammatical Subgrouping of Formosan Languages. In Jen kuei Li, Cheng hwa Tsang, Ying kuei Huang, Dah an Ho, and Chiu yu Tseng, editors, *Austronesian Studies Relating to Taiwan*, pages 683–726. Institute of History and Philology, Academia Sinica, Taipei.

Elizabeth Zeitoun. 2007. *A Grammar of Mantauran (Rukai)*. Institute of Linguistics, Academia Sinica, Taipei.

Elizabeth Zeitoun and Yoann Goudin. 2024. Language Contact in Formosan Languages. *Handbook of Formosan Languages: The Indigenous Languages of Taiwan*, 2:246–283.

Elizabeth Zeitoun, Ching-hua Yu, and Cui-xia Weng. 2003. The Formosan Language Archive: Development of a Multimedia Tool to Salvage the Languages and Oral Traditions of the Indigenous Tribes of Taiwan. *Oceanic Linguistics*, 42(1):218–232.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving Low-Resource Machine Translation for Formosan Languages Using Bilingual Lexical Resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11248–11259, Bangkok, Thailand. Association for Computational Linguistics.

9. Language Resource References

Indigenous Languages Research and Development Foundation. 2020. 族語 *E* 樂園. Accessed: 2026-01-17.

Mohamed, W. and Le Ferrand, É. and Sung, L.-M. and Prud'hommeaux, E. and Hartshorne, J. K. 2024. *FormosanBank*. Electronic Resource; access date: 2026-01-17.

A. Dendrograms

We present dendrograms based on TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance in Figure 2.

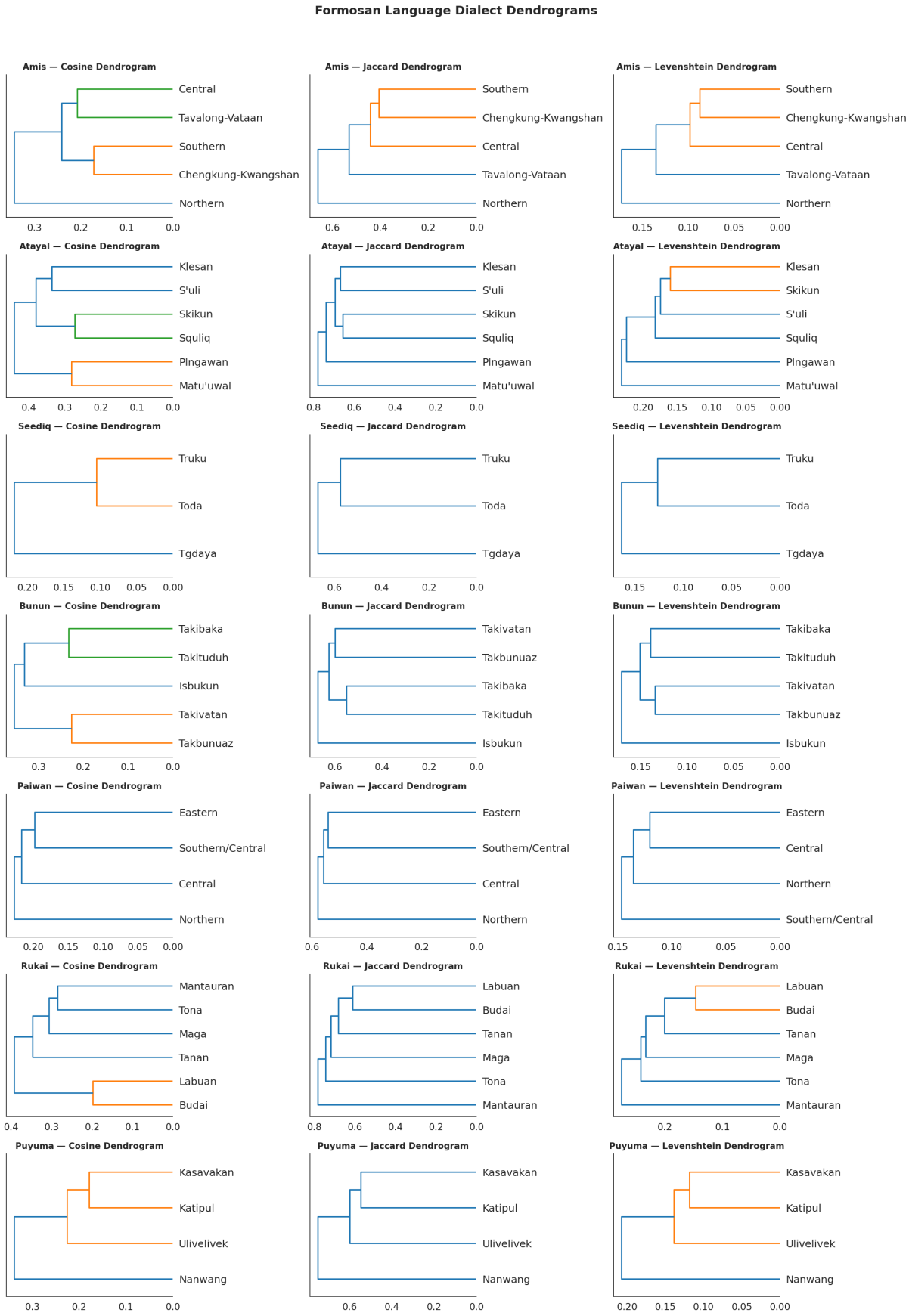


Figure 2: Per-dialect dendrograms based on TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance for all seven Formosan languages with multiple dialects

A Dialectal Corpus for Ukrainian: Collection, Classification, and Standardization

Yuliia Frund **Sina Ahmadi**
Department of Computational Linguistics
University of Zurich
{yuliia.frund,sina.ahmadi}@uzh.ch

Abstract

Ukrainian dialects remain largely excluded from the digital linguistic landscape despite their active everyday use. We present a regional dialect corpus covering 18 administrative regions of Ukraine, compiled from digitized fieldwork collections and an online dialect atlas. The corpus comprises over 284,000 tokens of dialect text, annotated by region and partially accompanied by manually standardized translations. Using these resources, we investigate language identification and dialect-to-standard standardization. Baseline language identification yields an F-score of 0.75, rising to 0.99 with dialect-inclusive training. Dialect classification reaches 0.58, with confusion patterns reflecting known regional boundaries. For standardization, the best-performing LLM achieves a COMET score of 0.80, though BLEU scores remain low (0.21–0.23) across all models. We release the corpus, labelled datasets, model outputs, and reference translations to support future work on inclusive language technologies for non-standard varieties. Our resource is openly available at https://github.com/yfrund/ukrainian_dialects.

Keywords: Ukrainian dialects, dialect corpus, language identification, dialect classification, low-resource NLP

1. Introduction

Dialects are actively used in everyday communication by millions of speakers worldwide, yet modern language technologies remain predominantly trained on standard language varieties. This creates a growing disconnect: as natural language processing (NLP) systems become more embedded in public services, healthcare, education, and digital communication, their inability to handle non-standard input excludes a significant portion of speakers. Recognizing this gap, recent years have seen increasing efforts to develop dialectal resources and tools across languages such as Arabic (Keleg et al., 2025), Greek (Chatzikyriakidis et al., 2023), Italian (Ramponi, 2024), and Germanic varieties (Borin, 2025), reflecting a broader shift toward inclusive language and speech technologies (Joshi et al., 2025).

Ukrainian is a language where this gap is particularly evident. While resources for Standard Ukrainian have expanded considerably, including large annotated corpora (Shvedova et al., 2017–2022), parliamentary proceedings (Kopp et al., 2023), and treebanks, its dialects remain largely absent from the digital landscape. Ukrainian encompasses three major dialect groups spanning the country's administrative regions (Del Gaudio, 2017), with substantial variation at the phonetic, morphological, syntactic, and lexical levels. For instance, Northern dialects exhibit distinctive diphthongization patterns and loss of consonant palatalization, South-western varieties preserve archaic morphological forms, and the South-eastern group, while closest to the standard variety, still displays notable phonetic and lexical divergence. This diversity is both a challenge for NLP systems and a valuable resource for building more robust tools, yet the lack of digitized, annotated dialect data has limited progress in this direction.

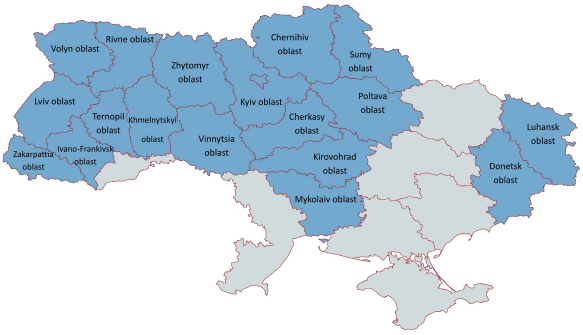


Figure 1: Regions in Ukraine for which our corpus contains dialectal data (highlighted in blue).

In this paper, we address this gap by collecting and curating dialectal data from digitized fieldwork collections and an online dialect atlas, covering 18 administrative regions of Ukraine with over 284,000 tokens. Using these resources, we investigate how current NLP tools handle dialectal input and whether they can be adapted to better accommodate non-standard variation. Our contributions are as follows:

- A dialectal corpus of Ukrainian annotated by region of origin;
- A dialect-to-standard reference set for machine translation, created through LLM-generated outputs with manual correction;
- Baseline experiments on language identification and dialect standardization.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 135–143
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

Dialectal NLP. The challenges of processing dialectal text have received growing attention in the NLP community. Joshi et al. (2025) provide a comprehensive survey of methods applied to dialects, identifying data scarcity, inconsistent orthography, and evaluation gaps as persistent bottlenecks. On the evaluation side, benchmarks such as CoDET (Alam et al., 2024) and DialectBench (Faisal et al., 2024) have been introduced to enable systematic assessment of NLP systems on dialectal data. Notably, Alam et al. (2024) report consistent performance drops in neural machine translation systems when processing dialectal input, further illustrating the impact of dialectal underrepresentation in training data. Language-specific efforts have addressed this gap through dedicated resource creation and evaluation for English (Ziems et al., 2023), Arabic (Keleg et al., 2025), Greek (Chatzikyriakidis et al., 2023), Italian (Ramponi, 2024), Kurdish (Ahmadi et al., 2024), and Germanic varieties (Borin, 2025; Plüss et al., 2023). A recurring finding across these efforts is that explicitly dialectal training data are often required to close the performance gap, as simply increasing data volume does not consistently help (Kantharuban et al., 2023).

Dialect normalization and standardization. A key challenge in dialect processing is that NLP tools are typically designed around standard orthographic conventions, making them ill-suited for dialectal text whose written forms do not conform to these norms. Prior work has explored rule-based (Shvedova et al., 2022), semi-automatic (Samardžić et al., 2016), and neural approaches (Kuparinen et al., 2023) to map non-standard text to standard forms, with transformer-based systems generally achieving the most robust results. However, these methods typically address only orthographic form, while dialects also differ in vocabulary and syntax. This raises the question of whether LLMs can standardize both form and vocabulary, bypassing the extensive annotation traditionally required, a possibility we investigate for Ukrainian dialects.

Resources for Ukrainian. Digital resources for Standard Ukrainian have grown in recent years. The GRAC corpus (Shvedova et al., 2017–2022) provides a large annotated collection of standard texts with advanced filtering capabilities. Kopp et al. (2023) introduced an annotated corpus of Ukrainian parliamentary proceedings. Community-driven projects such as Lang-UK[1] and the Universal Dependencies treebank for Ukrainian[2] further contribute to the standard-language resource ecosystem. However, these resources focus exclusively on Standard Ukrainian or on a specific dialect, as for Hutsul (Kyslyi et al., 2025). Dialectal text remains underrepresented, and to the best of our knowledge, no publicly available parallel corpus for Ukrainian dialects existed prior to this work.

3. Ukrainian Dialects

Ukrainian dialects are traditionally grouped into three major categories: Northern, South-western, and South-eastern (Del Gaudio, 2017). The Northern group is spoken across the Volyn, Rivne, Zhytomyr, Kyiv, Chernihiv, and Sumy regions. The South-western group spans from the western border to the west of Kyiv, Cherkasy, Kirovohrad, and Mykolaiv regions. The South-eastern group covers the remaining regions, from Kyiv and Chernihiv in the north to Odesa and Crimea in the south. Our corpus covers 18 of Ukraine's administrative regions across all three groups (see Figure 1).

The **Northern group** is phonetically the most distinct. Notable features include the development of etymological [o] into diphthongs in closed syllables, the pronunciation of unstressed [o] closer to [a], and the loss of palatalization in certain consonants; for example, standard буряк [burʲak] ('beetroot') becomes бурак [burak]. The group also exhibits morphosyntactic particularities, such as non-standard future-tense constructions (му ходити alongside standard буду ходити 'I will be going') and special numeral agreement patterns.

The **South-western group** displays significant local variation and preserves archaic forms. The Zakarpattia dialect retains што ('what'), and the reflexive particle ся functions as a clitic rather than the verbal suffix found in the standard variety. Other features include variation in demonstrative pronouns (e.g., сей/цей for 'this'), long and short possessive pronoun forms (моему/мойому 'mine', dative), and consonant hardening in certain noun declensions.

The **South-eastern group**, whose core lies in the Cherkasy and Poltava regions, forms the basis of modern Standard Ukrainian and is the most homogeneous of the three. Nevertheless, it displays features such as the approximation of [o] towards [u], realization of /f/ as [xv] (e.g., хвабрика instead of фабрика 'factory'), and parallel infinitive forms (робити/робить 'to do') (Del Gaudio, 2017).

This variation poses considerable challenges for language technologies trained exclusively on standard text, and motivates the broad geographical coverage of our corpus.

[1] https://lang.org.ua/en

[2] https://universaldependencies.org/uk

Name	Sources	Metadata
Dialect Map	Dialect terms added by philology students from various linguistic resources.	Part of speech, source, synonyms, use in context, set phrases
Hibeba and Lesnova (2019)	Recorded by field linguists, school teachers, students, language enthusiasts, and researchers.	Oblast, place of residence, speaker's name and birth year, date of recording
Martynova (2012)	Recorded by the compiler.	Place of residence, speaker's name, birth year, education
Martynova et al. (2013)	Recorded by the compilers and dialectology students.	Place of residence, speaker's name, birth year, education
Lesnova (2013)	Recorded by the compiler.	Place of residence, speaker's name and birth year
Kovalenko and Kovalenko (2019)	Recorded by the compilers.	Place of residence, speaker's name, birth year and education

Table 1: Overview of the resources used, along with their sources and additional metadata.

4. Methodology

4.1. Data Collection

Due to the lack of large-scale dialect corpora for Ukrainian, especially those annotated by region, we compiled a dataset from several publicly available resources. While dialect text can also be found on social media or in local publications, the former is noisy due to slang and inconsistent spelling, and the latter tends to focus on isolated word lists. We therefore limited our sources to online platforms with contextualized examples and digitized academic collections of connected dialect text:

- *Dialect Map*,[3] an interactive atlas providing usage examples across seven regions (Lviv, Ivano-Frankivsk, Volyn, Zakarpattia, Mykolaiv, Rivne, and Vinnytsia);
- Hibeba and Lesnova (2019), dialectal speech transcriptions covering 17 oblasts in which speakers describe the preparation of borshch—a culturally significant dish chosen to elicit naturalistic, topic-controlled speech across regions;
- Martynova (2012) and Martynova et al. (2013), dialect texts from Poltava and Cherkasy regions respectively;
- Lesnova (2013) and Kovalenko and Kovalenko (2019), complementary texts from Luhansk and Khmelnytskyi oblasts.

Table 1 provides an overview of these resources and the metadata they supply.

The raw data required substantial preprocessing. The web-based data were parsed from HTML, while the remaining collections were extracted from digitized PDFs. During filtering, only dialect content was retained; speaker metadata and interviewer questions in the standard variety were excluded. Since the sources represent dialect speech using phonetic transcription, we removed diacritics and phonetic symbols, converting non-standard characters to their closest standard equivalents (e.g., non-syllabic [ў] and [ĭ] to у and і). The raw data are structured into phonetic phrases delimited by pauses: short pauses were mapped to commas and long pauses to sentence boundaries as follows:

Raw Text
но ў на́шўў с'імйі́ шчи йе йиеде́н рецепт боуршчу́ // його́ варит мо́а ма́ма лиеше́ на с'а́ткы //
Cleaned Data
1. но у нашуу сімйі шчи йе йиден рецепт боршчу 2. його варит моа мама лише на саткы

Table 2: Raw vs. cleaned data ("But in our family there is one more recipe of borshch. My mother cooks it only for celebrations."). Non-standard tokens are underlined.

Another challenge was character encoding inconsistency in PDF-extracted text, where visually identical Latin and Cyrillic characters (e.g., *i*, *e*, *a*) had different Unicode values, affecting both statistics and model training. All Latin characters were mapped to their Cyrillic equivalents. For the purposes of this work, a dialect is defined as corresponding to an administrative region (oblast), yielding 18 classes. While this does not capture the full variation within each region, it provides a practical framework that balances linguistic relevance with

[3] https://dialectmap.org/en

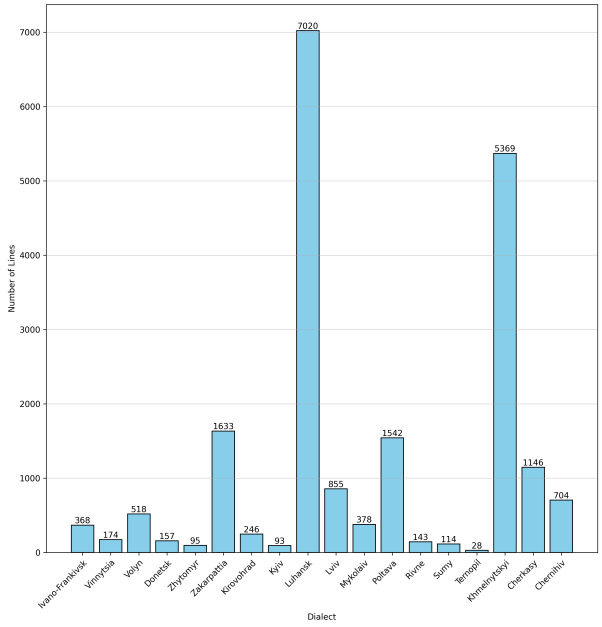


Figure 2: Class distribution of dialect data by oblast. Luhansk is the largest class and Ternopil the smallest.

Metric	Value
Number of dialects	18
Number of lines	20,583
Number of tokens	284,045
Number of unique tokens	87,510
Avg. tokens per line	13
Avg. unique tokens per line	4
Avg. token length	5

Table 3: General corpus statistics.

scalability. Table 3 presents overall corpus statistics and Figure 2 shows the class distribution.

4.2. Reference Translations

To evaluate how well models can translate dialectal text to other languages and even Standard Ukrainian, we need parallel data pairing dialect input with its standard equivalent. Since no such resource exists for Ukrainian dialects, we create one semi-automatically by generating translations with multiple LLMs and then manually correcting them.

For each of the 18 dialect classes, 100 sentences were sampled (with a minimum length of 5 tokens to ensure sufficient context); for smaller classes, as many sentences as possible were taken and the remainder supplemented from larger classes, yielding 1,800 sentence pairs in total. Four models were used to generate initial translations: three quantized models (EuroLLM-9B-Instruct (Martins et al., 2025), Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.1 (Jiang et al., 2023)), along with GPT-3.5-turbo. All models received the prompt: *"Translate the text into Standard Ukrainian. Do not provide any commentary."* with a temperature of 0 to ensure reproducibility.

To avoid biasing the reference set toward any single model's outputs, translations were sampled in a round-robin manner: for each source sentence, the reference was taken from a different model in turn. The sampled translations were then manually reviewed and corrected with close attention to the source material. Since the data represent transcribed speech, sentence structure was preserved to avoid favouring any particular model's rephrasing style, and dialect-specific terms were retained when their status as non-standard was uncertain. Additionally, a set of translations was generated using NLLB (Team, 2024), but these were excluded from the reference set and evaluated separately.

4.3. Experimental Setup

We conduct experiments on two tasks: language identification and dialect-to-standard text standardization.

Language identification. We use fastText (Joulin et al., 2017, 2016) for language identification, as it supports 176 languages and can be trained efficiently on a standard CPU. Our baseline is the pretrained `lid.176.bin` model. We evaluate it by mapping all dialect text to a single Ukrainian label to assess how well dialect input is recognized as Ukrainian.

We then train two additional models. The *macro model* is trained on dialect data labelled as Ukrainian alongside samples from languages that caused frequent misclassifications in the baseline (Bulgarian, Macedonian, Serbian, Tatar, Russian, and Belarusian), sourced from the Leipzig Corpora Collection (Goldhahn et al., 2012). The *micro model* is trained on the same data but with dialect-specific labels (one per oblast), enabling inter-dialect classification. Since fastText does not support fine-tuning, pretrained word vectors were used to initialize representations for the macro model; the micro model was trained from scratch. Given the uneven class distribution, downsampling and upsampling were applied to balance classes at 500 samples each, with an 80/20 train/test split. Upsampling was performed after splitting to prevent data leakage. All models are evaluated using macro F-score. Table 4 reports the final hyperparameters.

Parameter	Macro Model	Micro Model
Vector dimensions	16	300
Learning rate	1.0	0.8
Epochs	60	80
Word n-grams	1	2
Min char n-gram	3	1
Max char n-gram	6	5
Pretrained vectors	Yes	No

Table 4: Final hyperparameter settings for fastText model training.

Dialect-to-standard standardization. We evaluate five systems on the task of mapping dialect text to Standard Ukrainian: three locally-run quantized LLMs (EuroLLM-9B-Instruct, Meta-Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.1), GPT-3.5-turbo, and the NLLB machine translation model (Team, 2024). The LLMs were quantized using GGUF formats (Q3_K_S, Q4_K_M, and Q4_K_S respectively) to enable inference on consumer-grade hardware (Lang et al., 2024). Translation quality is assessed using two complementary metrics: BLEU (Papineni et al., 2002), which measures surface-level n-gram overlap, and COMET (Rei et al., 2020), which evaluates semantic adequacy. Each model receives a dialectal sentence as input and is prompted to produce its Standard Ukrainian equivalent. The generated output is then compared against the manually corrected reference translations using BLEU and COMET. We additionally test a more explicit prompt—*"Translate the text into Standard Ukrainian. Use standard spelling and standard vocabulary. Do not provide any commentary."*—to investigate prompt sensitivity. Scores are reported per dialect and in aggregate.

5. Experiments

5.1. Language Identification

The baseline fastText model achieves an overall macro F-score of 0.75 when classifying dialect text as Ukrainian. Performance varies by region, ranging from 0.56 (Sumy) to 0.99 (Mykolaiv), with per-dialect scores reported in Table 5. Misclassifications stem primarily from confusion with other East and South Slavic languages—particularly Russian, Belarusian, Bulgarian, Serbian, Macedonian, and Tatar—as shown in Figure 3.

The macro model, trained with dialect data labelled as Ukrainian alongside confusable languages, raises the F-score to 0.99, correctly identifying virtually all dialect input as Ukrainian (Figure 4). On the other hand, the micro model, trained with dialect-specific labels, achieves an F-score of 0.58. Figure 5 reveals several clusters of inter-dialect confusion. A western cluster groups Zakarpattia, Lviv, and Volyn, while a central-eastern cluster links Cherkasy, Poltava, Kirovohrad, and Chernihiv. Some dialect pairs show symmetric confusion (e.g., Poltava–Cherkasy: 29 and 24 misclassifications respectively; Luhansk–Khmelnytskyi: 15 and 10), while others are asymmetric (e.g., Mykolaiv misclassified as Lviv 21 times, but Lviv as Mykolaiv only 5 times).

To investigate these confusion patterns, we computed the top-10 most frequent bigrams through five-grams for each dialect and measured pairwise Jaccard similarity (Jaccard, 1901). As Figure 6 shows, confusion patterns largely align with n-gram similarity: Cherkasy and Poltava share a similarity of 0.86, Lviv and Volyn 0.57, and Zakarpattia and Lviv 0.48. The unexpected Luhansk–

Dialect	F	Dialect	F
Cherkasy	0.86	Mykolaiv	0.99
Chernihiv	0.63	Poltava	0.89
Donetsk	0.67	Rivne	0.64
Ivano-Frankivsk	0.79	Sumy	0.56
Khmelnytskyi	0.86	Ternopil	0.71
Kirovohrad	0.75	Vinnytsia	0.79
Kyiv	0.70	Volyn	0.65
Lviv	0.73	Zakarpattia	0.65
Luhansk	0.78	Zhytomyr	0.81

Table 5: Baseline F-scores: performance varies widely across dialects, from 0.56 (Sumy) to 0.99 (Mykolaiv).

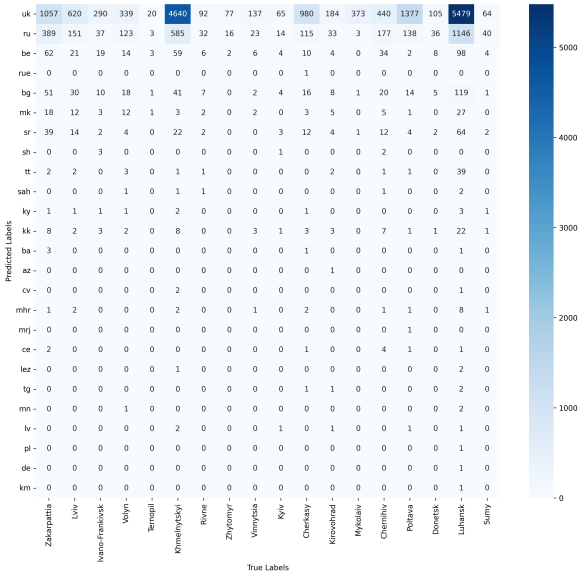


Figure 3: Baseline confusion matrix: Ukrainian dialect text is frequently misclassified as related Slavic languages. Dialects ordered West to East.

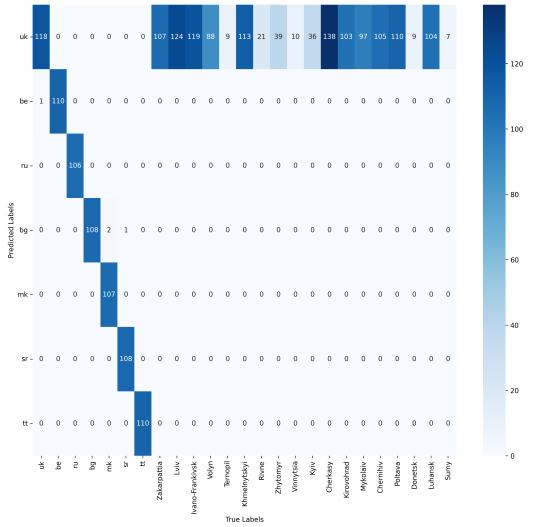


Figure 4: Macro model confusion matrix: including dialect data in training nearly eliminates cross-language misclassification (F=0.99).

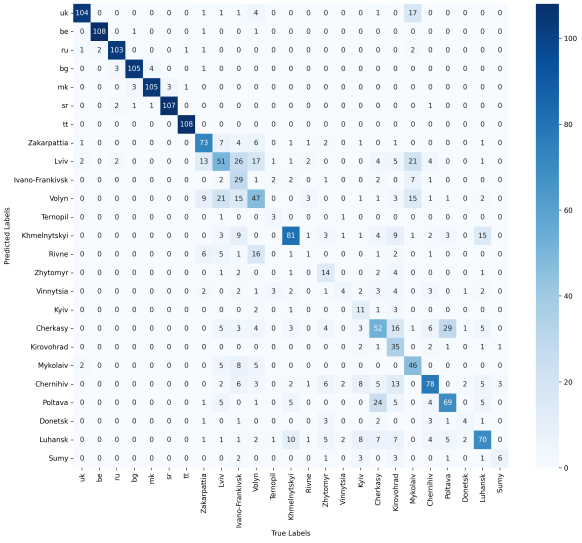


Figure 5: Micro model confusion matrix: inter-dialect confusion clusters emerge along geographic lines (F=0.58).

Khmelnytskyi confusion is also explained by their having the highest pairwise similarity among all non-neighbouring dialect pairs. However, high similarity does not always produce frequent misclassifications: Chernihiv and Sumy share a score of 0.54 yet yield only three misclassifications, suggesting the model can learn to distinguish similar dialects when sufficient training data are available.

5.2. Dialect-to-Standard Standardization

Table 6 presents BLEU scores across all models and dialects. Overall performance is low, with

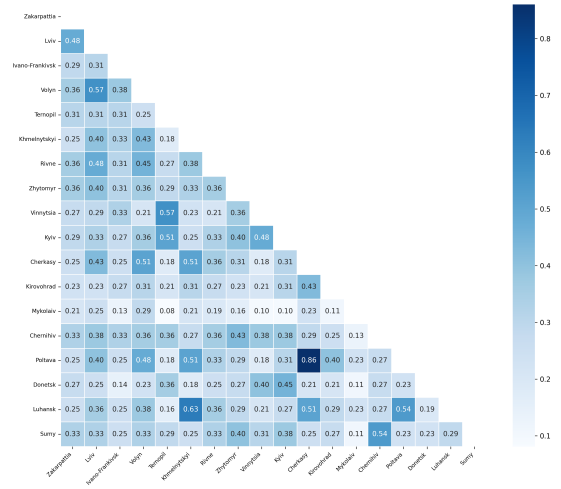


Figure 6: Jaccard similarity between dialect pairs based on n-gram overlap: geographically adjacent varieties tend to share more in common (e.g., Cherkasy–Poltava: 0.86, Lviv–Volyn: 0.57).

scores ranging from 0.21 (LLaMA, NLLB) to 0.23 (EuroLLM). All models consistently score lowest on Donetsk (0.07) and highest on Kyiv (0.36–0.45). The four LLMs perform comparably, while NLLB trails slightly. When it comes to COMET (Table 7), GPT-3.5 leads with 0.80, followed by EuroLLM (0.76), LLaMA (0.73), Mistral (0.66), and NLLB (0.56). All models perform worst on Sumy and best on Mykolaiv. The divergence between BLEU and COMET rankings indicates that while surface-level overlap with the reference remains limited, models, particularly GPT-3.5 and EuroLLM, preserve semantic content well.

Manual inspection reveals qualitative differences across models. GPT-3.5 produces the most consistent output in terms of capitalization and punctuation. EuroLLM occasionally enters infinite generation loops and sometimes conflates the system prompt with the output. Mistral shows issues with diacritics and inconsistent script, sometimes producing Latin-script transliterations instead of Cyrillic. LLaMA occasionally refuses to translate, misinterpreting dialectal input as incomplete or misleading. Note that NLLB translations were generated after the reference set was finalized and are therefore not included in it, placing NLLB at a disadvantage in direct comparison.

A more explicit prompt (*"Use standard spelling and standard vocabulary."*) produced only minor score variations, with the overall model ranking unchanged. The most notable shifts were observed for Vinnytsia (BLEU: 0.32→0.25) and Rivne (BLEU: 0.36→0.23).

Dialect	EuroLLM	GPT-3.5	LLaMa	Mistral	NLLB
Ivano-Frankivsk	0.19	0.17	0.17	0.18	0.16
Vinnytsia	0.33	0.32	0.32	0.34	0.30
Volyn	0.13	0.13	0.13	0.13	0.06
Donetsk	**0.07**	**0.07**	**0.07**	**0.07**	**0.06**
Zhytomyr	0.22	0.18	0.21	0.23	0.18
Zakarpattia	0.21	0.21	0.15	0.21	0.19
Kyiv	**0.45**	**0.41**	**0.39**	**0.39**	**0.36**
Kirovohrad	0.29	0.27	0.25	0.28	0.25
Luhansk	0.10	0.10	0.10	0.10	0.09
Lviv	0.38	0.36	0.31	0.36	0.31
Mykolaiv	0.22	0.22	0.21	0.22	0.22
Poltava	0.27	0.26	0.25	0.27	0.23
Rivne	0.34	0.36	0.32	0.34	0.32
Sumy	0.25	0.26	0.26	0.26	0.12
Ternopil	0.14	0.13	0.13	0.14	0.13
Khmelnytskyi	0.08	0.12	0.09	0.11	0.11
Cherkasy	0.15	0.15	0.14	0.14	0.14
Chernihiv	0.32	0.32	0.30	0.34	0.28
Total	**0.23**	**0.22**	**0.21**	**0.22**	**0.21**

Table 6: Performance of Ukrainian dialect-to-standard translation based on BLEU.

Dialect	EuroLLM	GPT-3.5	LLaMa	Mistral	NLLB
Ivano-Frankivsk	0.75	0.80	0.71	0.68	0.56
Vinnytsia	0.78	0.84	0.76	0.70	0.57
Volyn	0.71	0.76	0.69	0.61	0.54
Donetsk	0.76	0.82	0.74	0.65	0.53
Zhytomyr	0.80	0.85	0.78	0.70	0.54
Zakarpattia	0.73	0.80	0.68	0.58	0.55
Kyiv	0.73	0.78	0.72	0.68	0.58
Kirovohrad	0.78	0.82	0.76	0.73	0.60
Luhansk	0.75	0.82	0.74	0.67	0.54
Lviv	0.74	0.81	0.72	0.66	0.58
Mykolaiv	**0.85**	**0.89**	**0.84**	**0.82**	**0.69**
Poltava	0.73	0.77	0.71	0.66	0.55
Rivne	0.71	0.74	0.66	0.61	0.55
Sumy	**0.65**	**0.70**	**0.62**	**0.57**	**0.50**
Ternopil	0.74	0.78	0.74	0.65	0.51
Khmelnytskyi	0.77	0.81	0.72	0.63	0.57
Cherkasy	0.84	0.85	0.78	0.73	0.58
Chernihiv	0.71	0.77	0.68	0.61	0.54
Total	**0.76**	**0.80**	**0.73**	**0.66**	**0.56**

Table 7: Performance of Ukrainian dialect-to-standard translation based on COMET.

6. Conclusion

We presented a dialectal corpus for Ukrainian covering 18 administrative regions, along with a dialect-to-standard parallel reference set built by combining LLM outputs with manual correction. We used these resources to run experiments on language identification and dialect-to-standard standardization. The results show that standard NLP tools struggle with dialectal input, a baseline fastText model scores only 0.75 on dialect text, but that adding even a modest amount of dialect data to training brings this up to 0.99. When we push the task further to distinguishing individual dialects, performance drops to 0.58, with confusion patterns that track geographic proximity and surface-level similarity between varieties. On the standardization side, GPT-3.5 achieves a COMET score of 0.80, which suggests that the semantic content is largely preserved, though low BLEU scores (0.21–0.23) across all models point to limited surface-level overlap with the reference. From a practical standpoint, running dialectal text through an LLM as a pre-processing step can make it more compatible with downstream tools.

Future Work Looking ahead, the most immediate need is more data: expanding the corpus to additional regions and domains would help address class imbalance and improve coverage. On the modelling side, transformer-based classifiers and few-shot or fine-tuned LLMs are natural next steps for both identification and standardization. Extending this line of work to speech would also be valuable, given that dialects are first and foremost spoken varieties.

Limitations Our reference translations were constructed by sampling LLM outputs in round-robin fashion and manually correcting them. While this ensures balanced model representation, the evaluation metrics, particularly BLEU, may still be influenced by shared output patterns across the generating models, as surface-level phrasing choices could carry over even after manual editing. A more principled selection strategy could mitigate this, for instance by leveraging quality signals from model outputs rather than fixed rotation. For scaling to a larger parallel corpus, active learning and LLM-as-a-judge approaches could help prioritize which sentences require human correction, reducing annotation effort while improving reference quality. Additionally, all corrections were performed by a single annotator without involvement of dialectologists, introducing potential subjectivity. Furthermore, the dialect boundaries used in our classification largely coincide with administrative borders, which do not necessarily reflect natural linguistic boundaries. Finally, the language identification models were trained on a small dataset of approximately 500 lines per class, with some minority classes containing as few as 28 lines, which limits the generalizability of classification results.

Ethics Statement All data were sourced from publicly available academic publications and an open online platform. Speaker metadata present in the original sources (names, birth years) is not redistributed; only textual content is used. We note that dialect classification could in principle be misused for regional profiling and encourage responsible use of the released resources.

Acknowledgments

Sina Ahmadi gratefully thanks the support of the UZH Grant (reference number 269093).

7. Bibliographical References

S. Ahmadi, D. Jaff, M. M. I. Alam, and A. Anastasopoulos. 2024. Language and speech technology for Central Kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045.

M. M. I. Alam, S. Ahmadi, and A. Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

L. Borin. 2025. Corpus and computational linguistic approaches to Germanic languages. In *Oxford Research Encyclopedia of Linguistics*.

S. Chatzikyriakidis, C. Qwaider, I. Kolokousis, C. Koula, D. Papadakis, and E. Sakellariou. 2023. Grdd: A dataset for Greek dialectal NLP. *arXiv preprint arXiv:2308.00802*.

S. Del Gaudio. 2017. *An Introduction to Ukrainian Dialectology*. Peter Lang Verlag, Berlin, Germany.

F. Faisal, O. Ahia, A. Srivastava, K. Ahuja, D. Chiang, Y. Tsvetkov, and A. Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

D. Goldhahn, T. Eckart, and U. Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

N. Hibeba and V. Lesnova, editors. 2019. *Treasures of Ukrainian Dialects: Texts about Borshch [Скарби українських говорів: тексти про борщ]*. Dialektolohichna skrynia. Instytut Ukrainoznavstva im. I. Krypiakevycha NAN Ukrainy, Lviv.

P. Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed. 2023. Mistral 7b.

A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, and D. Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

A. Kantharuban, I. Vulić, and A. Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

A. Keleg, S. Goldwater, and W. Magdy. 2025. Revisiting common assumptions about Arabic dialects in NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3327.

M. Kopp, A. Kryvenko, and A. Rii. 2023. Ukrainian parliamentary corpus ParlaMint-UA 4.0.1. Slovenian language resource repository CLARIN.SI.

N. Kovalenko and B. Kovalenko, editors. 2019. *Volhynian Dialects of Khmelnychchyna: A Collection of Dialect Texts [Волинські говірки Хмельниччини. Збірник діалектних текстів]*. TOV "Ruta", Kamyanets-Podilskyi.

O. Kuparinen, A. Miletić, and Y. Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 13814–13828, Singapore. Association for Computational Linguistics.

Roman Kyslyi, Yuliia Maksymiuk, and Ihor Pysmennyi. 2025. Vuyko Mistral: Adapting LLMs for low-resource dialectal translation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 86–95, Vienna, Austria (online). Association for Computational Linguistics.

J. Lang, Z. Guo, and S. Huang. 2024. A comprehensive study on quantization techniques for large language models. In *2024 4th International conference on artificial intelligence, robotics, and communication (ICAIRC)*, pages 224–231. IEEE.

V. Lesnova. 2013. *Dialects of Eastern Slobozhanshchyna: A Collection of Dialect Texts [Говірки Східної Слобожанщини: збірник діалектних текстів]*. Vyd-vo DNU imeni Tarasa Shevchenka, Luhansk.

P. H. Martins, J. Alves, P. Fernandes, N. M. Guerreiro, R. Rei, A. Farajian, M. Klimaszewski, D. M. Alves, J. Pombal, N. Boizard, et al. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

H. Martynova. 2012. *Dialects of Western Poltava Region: A Collection of Dialect Texts [Говірки Західної Полтавщини: зб. діалектних текстів]*. Bohdan Khmelnytsky National University of Cherkasy, Cherkasy.

H. Martynova, T. Shcherbyna, and A. Taran. 2013. *Dialects of the Cherkasy Region: A Collection of Dialect Texts [Говірки Черкащини: збірник діалектних текстів]*. Bohdan Khmelnytsky National University of Cherkasy, Cherkasy.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

M. Plüss, J. Deriu, Y. Schraner, C. Paonessa, J. Hartmann, L. Schmidt, C. Scheller, M. Hürlimann, T. Samardžić, M. Vogel, and M. Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

A. Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

R. Rei, C. Stewart, A. C Farinha, and A. Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

T. Samardžić, Y. Scherrer, and E. Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

M. Shvedova, N. Prydvorova, and I. Skibina. 2022. Normalization of early modern Ukrainian in GRAC: The case of Lesia Ukrainka's works. In *COLINS*, pages 71–80.

M. Shvedova, R. von Waldenfels, S. Yarygin, A. Rysin, V. Starko, T. Nikolajenko, and others. 2017–2022. GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource: Kyiv, Lviv, Jena. Available at `http://uacorpus.org`.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

C. Ziems, W. Held, J. Yang, J. Dhamala, R. Gupta, and D. Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768.

German Dialects Across Situations, Generations, and Regions: The REDE corpus as an Oral Resource for NLP

Hanna Fischer, Alfred Lameli
Research Center Deutscher Sprachatlas, Marburg University
Pilgrimstein 16, 35032 Marburg, Germany
{hanna.fischer, lameli}@uni-marburg.de

Abstract

Recent advances in speech and language technologies increasingly rely on large and diverse corpora that represent linguistic variation across dialect regions, communicative situations, and social speaker characteristics. While substantial resources are available for Standard German, comparable spoken corpora for German dialects have so far been largely lacking, limiting the development and evaluation of dialect-sensitive NLP systems. The REDE corpus addresses this gap by providing a methodologically uniform collection of spoken German for 148 locations that systematically covers all major dialect areas in Germany. It comprises contemporary recordings collected in multiple elicitation and interaction settings, capturing variation across speaking styles, situational contexts, and speaker generations. With more than 1,500 hours of speech and rich metadata on regional and social dimensions, the REDE corpus constitutes a large-scale oral resource suitable for both linguistic research and NLP applications. This paper presents the design, structure, and methodological foundations of the corpus and discusses its relevance for current speech technology requirements.

Keywords: corpus development, speech recordings, German, dialects, regional accents, REDE corpus, Wenker sentences

1. Introduction

The current state of the art in natural language processing (NLP) and regional language research—particularly with regard to spoken language data—is characterized by a marked discrepancy between rapid advances in applications for standard languages and persistent challenges in the acquisition, modeling, and processing of regional varieties (Netzwerk Regionale Sprache und Künstliche Intelligenz, 2026). Transformer-based models such as BERT (Devlin et al., 2019) and self-supervised speech representations such as wav2vec 2.0 (Baevski et al., 2020) have dramatically improved performance on Standard German tasks, and generative systems such as GPT (OpenAI et al., 2024) have achieved widespread popularity. This development is attributable to the availability of extensive corpora of both spoken and written Standard German (Kupietz and Schmidt, 2018; Deppermann et al., 2023).

Within this context, automatic dialect identification (ADI) emerges as a central task. ADI not only enables new insights into the characteristics, distribution, and evolution of dialects, but also provides the necessary foundation for deploying dialect-sensitive AI systems—such as automatic speech recognition (ASR) or advanced generative language models—by systematically accounting for dialectal variation (Yan and Vaseghi, 2002). However, for dialects and minority languages, such functional language models are still largely lacking. The primary reason is the absence of suitable training data: for most regional varieties, there are insufficient spoken and written resources available.

This paper focuses on a corpus of regional varieties of German (dialects, regional accents), where this problem also applies. While ASR systems for Standard German—such as the Whisper model (Radford et al., 2023)—and speech synthesis technologies like the Eleven Multilingual v2 model (ElevenLabs[1]) have already achieved high levels of accuracy, comparable systems for dialects and other regional varieties remain either largely unavailable or perform significantly worse (Gilles et al., 2023; Dolev et al., 2024; Fischer et al., 2025).

What is therefore required is a corpus that meets the technical requirements of NLP methods while adequately representing regional diversity. Such a corpus must satisfy several requirements: (i) regionally diversified coverage of different dialect areas; (ii) it must include contemporary recordings that reflect present-day linguistic realities; (iii) recordings must be collected using a uniform methodology to ensure comparability and high recording quality; (iv) the corpus should encompass different speaking styles and modes to capture real-world variation—both in terms of dialect intensity (regional accents vs. local dialects) and experimental settings (e.g., controlled tasks vs. free speech); (v) to reflect social differentiation, the corpus should include speakers from different generations; (vi) the availability of transcriptions (orthographic, phonetic, or conversation-analytic) is required for a range of

[1] https://elevenlabs.io

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 144–152
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

analytic approaches; (vii) legal requirements for usability of the data must be met and the recordings must be technically prepared and accessible in open standard file formats.

The REDE corpus is the first nationwide, methodologically uniform spoken language resource for German dialects that meets all seven of these requirements. The aim of this paper is to present the design of the corpus, its structure, and relevance for linguistic research and NLP.

2. The REDE corpus

2.1. Context and Objectives

The REDE corpus was developed within the still ongoing research project Regionalsprache.de (REDE) (Schmidt et al., 2020–). One of the project's core objectives is the first systematic collection, analysis, and documentation of contemporary regional languages of German. The concept of *regional languages* encompasses all speech forms within the variation space between local dialects and regionally accented Standard German (Kehrein, 2019b; Schmidt and Herrgen, 2011).

2.2. Data Collection

The speech recordings were collected between 2008 and 2015 following a uniform methodological protocol. Data collection was conducted by twelve trained fieldworkers at a total of 148 locations across Germany. Location selection aimed at balanced regional coverage while also prioritizing sites that functioned as emergency call centers, as the core speaker group consisted of police officers engaged in emergency call handling. (For an overview of all locations and their assignment to the dialect areas, see the maps in in Figure 1 and Figure 2. For a tabular overview of the number of locations, speakers, and recordings per dialect area, see Table 1.

Recordings were made in stereo using high-quality microphones and recorders (microphone: *SONY ECM-MS957*, recorder: *MARANTZ professional PMD661 handheld solid state recorder*) at a sampling rate of 44.1 kHz and a bit depth of 16 bit. All sessions were conducted in quiet, controlled environments (e.g., offices, meeting rooms) at the respective police stations or fieldwork locations.

2.3. Speakers

Speech data were collected from three speaker groups differing in age and occupation (see Table 2). Group 2 (G2) constitutes the core group: male middle-aged police officers with comparable educational backgrounds and professional profiles,



Figure 1: Location network of the REDE corpus (see REDE-Infothek (Lipfert, 2024b) for location codes and Figure 2 for color coding).

who also participated as speakers in REDE's predecessor project, the so-called DiGS project[2].

In addition, speakers from two comparison groups were recorded that differ from the police officers with respect to age, educational background, and profession (see Table 2). For all REDE speaker groups, the selection criterion of local rootedness—at least second, ideally third generation—applied. In total, recordings were made from 780 speakers (all male; see Section 5 for a discussion of this limitation and its implications for NLP applications).

2.4. Recording Methodology

The recordings were conducted on site in direct interaction with the speakers. Each speaker was recorded in up to five recording situations that vary with respect to speech mode (translation, reading, conversation) and intended variety (local dialect vs. Standard German). In addition to three controlled settings, two free speech settings were included (see Table 3). The recording situations can be characterized as follows.

In the WS-D setting, speakers were asked to translate standard-language Wenker sentences[3]

[2]The DiGS project was conducted in cooperation with the German Federal Criminal Police Office (BKA).

[3]Wenker sentences are a standardized set of 40 elicitation sentences that have been widely used in dialect sur-

Dialect macro regions	Dialect regions (with transition zones)	Locations	Speakers	Recordings
Low German				
	Northern Low German	13	70	343
	N. Low German-Mecklenb.-West Pom.	1	5	25
	Mecklenburgish-West Pomeranian	9	43	212
	Mecklenb.-West Pom.-Brandenburgish	1	4	20
	Central Pomeranian	1	6	30
	Westphalian	8	49	240
	Eastphalian	4	23	112
	Eastphalian-Brandenburgish	1	6	28
	Eastphalian-Northern Low German	1	5	25
	Brandenburgish	6	27	132
	Brandenburgish-South Markish	2	10	50
Subtotal Low German		47	248	1217
West German				
	Low Franconian	2	9	45
	Ripuarian-Low Franconian	3	16	78
	Ripuarian	5	25	117
	Ripuarian-Moselle Franconian	2	11	55
	Moselle Franconian	7	42	203
	Moselle Franconian-Rhine Franconian	2	11	52
	Moselle Franconian-Central Hessian	2	9	45
Subtotal West German		23	123	595
Central German				
	Berlin	1	4	20
	North Upper-Saxon-Low Markish	2	8	40
	Upper Saxon	4	19	78
	Thuringian	4	21	104
	Thuringian-Upper Saxon	5	23	112
	Silesian	1	6	24
	North Hessian	3	15	74
	East Hessian	1	12	59
	Central Hessian	5	30	140
	Central Hessian-North Hessian	1	4	20
	Central Hessian-Rhine Franconian	1	8	37
	Rhine Franconian	9	46	226
	Rhine Franc.-Mosl. Franconian	1	4	19
	Rhine Franc.-Mosl. Franc.-Central Hessian	1	5	25
	Rhine Franc.-Low Alemannic	1	5	24
Subtotal Centr. German		40	210	1002
Upper German				
	Low Alemannic	2	8	40
	Central Alemannic	3	13	65
	Swabian	9	48	234
	Swabian-East Franconian	1	6	28
	High Alemannic	1	5	25
	High Alemannic-Low Alemannic	1	4	20
	East Franconian	8	42	203
	East Franconian-North Bavarian	1	5	25
	North Bavarian	3	14	70
	North Bavarian-Central Bavarian	1	6	30
	Central Bavarian-North Bavarian	1	7	33
	Central Bavarian	5	29	144
	Central Bavarian-Swabian-South Bavarian	1	7	34
	Bavarian-Swabian	1	5	22
Subtotal Upper German		38	199	973
Total		**148**	**780**	**3787**

Table 1: Overview of locations, speakers, and recordings by dialect area; dialect macro regions following Lameli (2013), subdivided according to Wiesinger (1983).

Group	Age range	Occupation	N
Group 1 (G1)	60+	manual occupations (e.g., crafts, agriculture)	260
Group 2 (G2)	42–59	police officers	326
Group 3 (G3)	17–26	upper secondary students and university students	194
Total			780

Table 2: Speakers



Figure 2: Classification of German dialects (Fischer and Lameli (forthcoming); Low German dialects are coloured in blue, Central German dialects in yellow, Upper German dialects in orange, and Western German dialects in green. Non-Germanic varieties, e.g., Sorbian, are not depicted on the map).

Abbreviation	Setting description
WS-D	Translation of Wenker sentences into the local dialect
WS-S	Translation of Wenker sentences into Standard German
NoSo	Reading task
I	Guided interview
FR	Conversation with a friend

Table 3: Recording settings

read aloud by the fieldworker into the local dialect. In WS-S, speakers listened to dialect recordings of Wenker sentences made prior to the survey that represented their own dialect region and translated them into Standard German. In the NoSo reading task, speakers read aloud Aesop's fable *Der Nordwind und die Sonne* provided in written form.

The interview (I) was conducted using a semi-structured guideline and recorded. Topics included speakers' language-biographical background, individual language use, language attitudes, and self-assessment of linguistic competence. In the friend conversation (FR) setting, a free conversation between the speaker and self-selected acquaintances from the same dialect region was recorded. The conversational partners were also recorded as part of the corpus; their speech is excluded from annotations. In addition, sociodemographic data were collected, and a questionnaire on linguistic background was completed.

Not all settings could be realized for every speaker, for example, because speakers did not understand the dialectal stimulus recordings in the WS-S translation setting. However, it was ensured that for each location and each speaker generation, at least one speaker was recorded in all settings.

2.5. Corpus Size and Structure

In total, the REDE corpus comprises 3,787 recordings with an overall duration of 1,542 hours. This corpus size enables fine-grained linguistic analyses and makes the resource attractive for NLP applications. Table 4 provides an overview of the number and total duration of recordings by setting. On average, interviews constitute the longest recordings (ca. 56 minutes), whereas the reading task lasts approximately one minute. Free speech settings (I and FR) together account for 72% of the total recording time, providing ample naturalistic data for language modeling and ASR. An overview on the number of available recordings per dialect area is provided by Table 1.

Setting	N	Total duration (hh:mm:ss)	Avg. duration (hh:mm:ss)
NoSo	769	12:15:34	00:00:57
WS-S	766	214:27:25	00:16:48
WS-D	770	212:21:41	00:16:33
I	765	711:09:29	00:55:47
FR	717	420:07:23	00:35:09
Total	3,787	1,570:21:32	00:24:53

Table 4: Number and duration of recordings by setting

veys since their original deployment in Georg Wenker's *Sprachatlas des Deutschen Reichs* (1889–1923) (see Lameli, 2014; Fleischer, 2017).

2.6. Data Processing and Documentation

The recordings are available as WAV files, organized by location, speaker, and recording setting. The speech recordings were further processed in specific subsamples. Orthographic and phonetic transcriptions are available as TextGrids for recordings from 86 locations—each including speakers from all three generations and all five situational contexts. For a further subsample of 60 locations, measures of dialect intensity were additionally performed and published within the REDE SprachGIS. Additional orthographic transcriptions exist for 1,000-word excerpts from 224 recordings from the FR and I settings for 56 locations. Phonetic transcriptions follow IPA conventions and are available at the phone, syllable and word level.

Furthermore, for FR of 56 young speakers (G3) at 48 locations, conversation-analytic transcriptions following the GAT2 conventions (Selting et al., 2009) were produced (Fischer et al., 2023).

Rich metadata are available for all recordings and speakers, including information on location, dialect area, language-biographical background, self-assessed dialect competence, and dialect use.

All documentation related to data collection and processing is provided in the REDE Infothek. Templates such as recording protocols and the linguistic background questionnaire, as well as data sets, are published in the LinguRep repository. A selection of results is available on the platform *Regionalakzente in Deutschland* (Pheiff and Pistor, 2023).

Currently, the audio recordings are available upon request and following the conclusion of a cooperation agreement. Publication in the LinguRep repository under a free license is in preparation.

3. The REDE corpus as a Research Resource

3.1. Suitability for Variationist Linguistics

Based on the REDE recordings, a wide range of qualitative and quantitative studies have investigated speakers' situational variation behavior. A particular focus has been placed on phonetic and phonological variation in individual regions of the language area covered by REDE (e.g., Kehrein, 2012, 2019a; Rocholl, 2015; Keil, 2017; Vorberger, 2019; Bohnert-Kraus, 2020; Stiel, 2020; Limper, 2024; Lameli, 2025). These studies employ measures of dialect intensity (measured as phonetic difference from codified spoken standard variety (Herrgen et al., 2001; Lipfert, 2024a; Lameli, forthcoming)) as well as qualitative analyses and statistical methods. Dialectal prosody has been examined by Pistor (2022, 2025a,b), while phonetic studies on vowel space (Kleen et al., 2024) and on pre-boundary lengthening (Spina and Lameli, 2024; Lameli and Spina, 2025; Spina and Lameli, 2025) demonstrate that the acoustic quality of the recordings is sufficient for fine-grained phonetic analyses. Morphological and morphosyntactic variation has been investigated by Fischer (2022) and Fischer and Rabanus (2023).

Taken together, these studies validate the acoustic and linguistic quality of the REDE corpus across multiple levels of analysis and confirm its suitability for systematic variationist linguistics.

3.2. Suitability for NLP Applications

Automatic dialect identification. The suitability of the REDE corpus for NLP applications has already been demonstrated in several studies. The recordings constitute a central data resource in the project *AnDy: Automatische Analyse der Dynamik dialektalen Sprechens mit Methoden der Künstlichen Intelligenz*. The project addresses both spatial and register-based classification of German dialects (see, e.g., Fischbach et al., 2025a,b) and evaluates speech processing methods such as speaker diarization (Fischbach, 2024).

Automatic speech recognition. Further NLP applications have drawn on REDE subsets: Bystrich (2023), for instance, used Bavarian Wenker sentences from REDE to evaluate ASR models. His work demonstrates that large corpora with orthographically transcribed data enriched with dialectal syntactic, lexical, and morphological features are essential for effectively combining standard language models with dialect-specific models. A similar objective was pursued by Lindner (2025), who adapted a standard ASR model to the Bavarian dialect using transfer learning on Wenker sentences recordings from three REDE locations. The REDE corpus has also been used in forensic applications: Siewert (2023) trained an ASR model on REDE conversational recordings and achieved highly accurate speaker localization.

Voice cloning and speaker localisation. In addition, Fischer et al. (2025) demonstrate that the REDE corpus can be used for modern voice cloning technologies, including speech synthesis systems such as the Eleven Multilingual v2 model (ElevenLabs), which are capable of replicating individual speaker voices with high fidelity (see also Gilles et al., 2023; Dolev et al., 2024).

4. Discussion

The REDE corpus is, to date, the only spoken language resource for German dialects that combines

nationwide coverage (locations across all dialect macro regions, i.e., Low German vs. High German), situational breadth (five recording settings), generational stratification, and a uniform collection protocol, yielding 1,542 hours with rich sociolinguistic metadata. While methodologically uniform corpora with dialect recordings do exist, these were collected in the 1950s and 1960s, for example, the Zwirner corpus (Zwirner, 1962) and the GDR corpus (Ehlers, 2022). Other corpora are restricted to specific dialect areas and therefore do not permit analyses of cross-dialect continua or convergence zones, for example, SwissCoco for Switzerland (Cieliebak et al., 2025), Schnëssen for Luxembourg (Entringer et al., 2021), DiÖ for Austria (Lenz, 2018), and SiN for northern Germany (Elmentaler et al., 2015). By contrast, regionally diversified corpora such as Deutsch heute corpus (Kleiner, 2015) focus exclusively on regionally accented reading pronunciation, while others, such as the Pfeffer corpus (Pfeffer, 1975), target regiolectal speech and exclude base dialects.

Through its methodological setup, the REDE corpus provides a curated, machine-readable resource for developing and evaluating AI systems. Its annotated structure supports both linguistic research and the development of language-aware AI systems that are responsive to regional variation. This is particularly relevant for accessible communication, human–machine interaction, and regionally adaptive voice technologies.

Furthermore, the REDE corpus complements existing resources such as DIALECTBENCH (Faisal et al., 2024) by its only spoken modality and also by providing evaluation data for underrepresented varieties and contributing to more inclusive benchmarking practices. It enables computational modeling of dialect continua and convergence zones, thereby opening new avenues for sociolinguistic and AI-driven analyses.

Beyond research applications, the REDE corpus is also available for dialect revitalization and preservation as well as for dialect studies and teaching. Central to these applications is the integration of the recordings into the linguistic geographic information system REDE SprachGIS, which provides cartographic access. In addition, REDE SprachGIS offers analyses of the dialect–standard continua for 60 REDE locations, each represented by one speaker from a given speaker group. These visualizations make regional language variation accessible for educational purposes (Pheiff et al., 2019; Schmidt, 2017).

5. Limitations

The REDE corpus has several limitations worth noting. First, coverage is restricted to Germany, excluding Austrian and Swiss dialect areas, though the methodologically parallel DiÖ corpus (Lenz, 2018) offers a complementary resource for Bavarian and Alemannic varieties. At present, there are no plans to extend the REDE corpus to additional German-speaking regions. Second, all speakers are male, reflecting the occupational gender bias of the core speaker group; this limits the transferability of NLP models trained on REDE data to female speech. Third, the recordings date from 2008–2015 and may not fully reflect current spoken varieties given ongoing dialect leveling. Finally, data annotation remains incomplete in parts and is being continuously expanded; access to the recordings is currently granted upon request and after conclusion of a cooperation agreement, with open publication in the LinguRep repository in preparation.

6. Acknowledgements

This research is supported by the *Academy of Sciences and Literature Mainz* (grant REDE 0404), the *Federal Ministry of Research, Technology and Space* (BMFTR) (grant AnDy 16DKWN007), and the Research Center *Deutscher Sprachatlas* of Marburg University. We are grateful to Lisa Dücker, Georg Oberdorfer, and Salome Lipfert for providing us with statistical overviews and information regarding the REDE data. We would also like to thank two anonymous reviewers for their valuable comments.

7. Bibliographical References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Mirja Bohnert-Kraus. 2020. *Regionalsprachliche Spektren im Mittelalemannischen*. Olms, Hildesheim.

Tobias Bystrich. 2023. Data-driven and rule-based approaches to improving Bavarian speech recognition. Bachelor's thesis, Universität Bonn and Fraunhofer-Institut IAIS.

Mark Cieliebak, Jonathan Gerber, and Manuela Hürlimann. 2025. SwissCoco2025 - the Swiss corpora collection 2025. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 133–148, Winterthur, Switzerland. Association for Computational Linguistics.

Arnulf Deppermann, Christian Fandrych, Marc Kupietz, and Thomas Schmidt, editors. 2023. *Korpora in der germanistischen Sprachwissenschaft*. De Gruyter, Berlin, Boston.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Eyal Liron Dolev, Clemens Fidel Lutz, and Noëmi Aepli. 2024. Does Whisper understand Swiss German? An automatic, qualitative, and human evaluation.

Klaas-Hinrich Ehlers. 2022. Die Tonbandaufnahmen der deutschen Mundarten im Kontext der (niederdeutschen) Dialektologie der DDR. IDSopen: Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache, No. 3.

Michael Elmentaler, Joachim Gessinger, Jens Lanwer, Peter Rosenberg, Ingrid Schröder, and Jan Wirrer. 2015. Sprachvariation in Norddeutschland (SiN). In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen: Projekte und Perspektiven*, pages 397–424. De Gruyter, Berlin and Munich and Boston.

Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke. 2021. Schnëssen. surveying language dynamics in luxembourgish with a mobile research app. *Linguistics Vanguard*, 7(s1).

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages.

Lea Fischbach. 2024. A Comparative Analysis of Speaker Diarization Models: Creating a Dataset for German Dialectal Speech. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 43–51, Bangkok. Association for Computational Linguistics.

Lea Fischbach, Akbar Karimi, Carolin Kleen, Alfred Lameli, and Lucie Flek. 2025a. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Proceedings of Interspeech 2025*, pages 2780–2784.

Lea Fischbach, Caroline Kleen, Lucie Flek, and Alfred Lameli. 2025b. Does Preprocessing Matter? An Analysis of Acoustic Feature Importance in Deep Learning for Dialect Classification. In *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–169, Tartu. University of Tartu Library.

Hanna Fischer. 2022. *Tempus und Regionalsprache. Eine gebrauchslinguistische Studie*. Olms, Hildesheim.

Hanna Fischer, Brigitte Ganswindt, and Georg Oberdorfer. 2023. Die regionalsprachlichen Tonkorpora des Forschungszentrums Deutscher Sprachatlas. In Marc Kupietz and Thomas Schmidt, editors, *Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022*, pages 189–202. Narr Francke Attempto, Tübingen.

Hanna Fischer and Alfred Lameli. forthcoming. Regional varieties of German. In Joshua Bousquette and Simon Pickl, editors, *The Oxford Handbook of the German Language*. Oxford University Press, Oxford.

Hanna Fischer, Alfred Lameli, Martha Schubert, and Ingo Siegert. 2025. Cloning dialects: Recreating and localizing dialectal voices. In *2025 IEEE International Professional Communication Conference (ProComm)*, pages 358–367.

Hanna Fischer and Stefan Rabanus. 2023. Zwischen dialektalem Hintergrund und standardsprachlicher Norm: verbalmorphologische Variation in standardintendierter Sprechweise. In Hanna Fischer and Stefan Rabanus, editors, *Morphologische und syntaktische Variation in den deutschen Regionalsprachen: Impulse für die Erforschung der sprachlichen Vertikale*. Olms, Hildesheim.

Jürg Fleischer. 2017. *Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen: Dokumentation, Entdeckungen und Neubewertungen*. Olms, Hildesheim and Zürich and New York.

Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023. LUX-ASR: Building an ASR system for the Luxembourgish language. In *Proceedings - 2022 IEEE Spoken Language Technology Workshop (SLT)*.

Joachim Herrgen, Alfred Lameli, Stefan Rabanus, and Jürgen Erich Schmidt. 2001. Dialektalität als phonetische Distanz. Ein Verfahren zur Messung standarddivergenter Sprechformen. Marburg University.

Roland Kehrein. 2012. *Regionalsprachliche Spektren im Raum – Zur linguistischen Struktur der Vertikale*. Steiner, Stuttgart.

Roland Kehrein. 2019a. 5. Areale Variation im Deutschen „vertikal“. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Volume 4 Deutsch*, pages 121–158. De Gruyter Mouton, Berlin, Boston.

Roland Kehrein. 2019b. Vertical language change in germany: Dialects, regiolects, and standard german. In Stanley D. Brunn and Roland Kehrein,

editors, *Handbook of the Changing World Language Map*, pages 1–13. Springer International Publishing, Cham.

Carsten Keil. 2017. *Der VokalJäger. Eine phonetisch-algorithmische Methode zur Vokaluntersuchung. Exemplarisch angewendet auf historische Tondokumente der Frankfurter Stadtmundart*. Olms, Hildesheim.

Caroline Kleen, Marina Frank, and Alfred Lameli. 2024. Vertical differences in German Vowel Space Areas. In *Proceedings of the 5th International Symposium on Applied Phonetics (ISAPh 2024)*, pages 43–48.

Stefan Kleiner. 2015. „Deutsch heute“ und der Atlas zur Aussprache des deutschen Gebrauchsstandards. In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen: Projekte und Perspektiven*, pages 489–518. De Gruyter, Berlin and Munich and Boston.

Marc Kupietz and Thomas Schmidt, editors. 2018. *Korpuslinguistik*. De Gruyter, Berlin, Boston.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. De Gruyter, Berlin and Boston.

Alfred Lameli. 2014. *Erläuterungen und Erschließungsmittel zu Georg Wenkers Schriften*. Olms, Hildesheim and New York and Zürich.

Alfred Lameli. 2025. Gesprochenes Deutsch in den Regionen. Eine Standortbestimmung für die Bundesrepublik Deutschland. In Nadine Proske, Thilo Weber, Monika Dannerer, and Arnulf Deppermann, editors, *Gesprochenes Deutsch. Struktur, Variation, Interaktion*, pages 51–79. De Gruyter, Berlin and Boston.

Alfred Lameli. forthcoming. Evaluating phonetically weighted and unweighted distance measures in dialectometry.

Alfred Lameli and Nadja Spina. 2025. The Signs of Time: The Indexical Value of Pre-boundary Lengthening. In Toke Hoffmeister, Christina Kauschke, and Mathias Scharinger, editors, *Repräsentationen aus linguistischer und interdisziplinärer Perspektive*, pages 357–381. Olms, Baden-Baden.

Alexandra N. Lenz. 2018. The special research programme: German in austria: Variation – contact – perception. *Sociolinguistica*, 32(1):269–278.

Juliane Limper. 2024. *Regionalsprachliche Spektren im Bairischen*. Steiner, Stuttgart.

Moritz Lindner. 2025. Anpassung eines Standard-Hochdeutsch-ASR-Modells an den bayerischen Dialekt mittels Transfer Learning. Master's thesis, Hochschule für Ökonomie & Management München.

Salome Lipfert. 2024a. REDE-Infothek: Phonetische Abstandsmessung (PAM). In Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer, editors, *Regionalsprache.de (REDE). Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Marburg.

Salome Lipfert. 2024b. REDE-Infothek: REDE-Orte. In Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer, editors, *Regionalsprache.de (REDE). Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Marburg.

Netzwerk Regionale Sprache und Künstliche Intelligenz. 2026. Regionale Sprache und Künstliche Intelligenz im Zeitalter der digitalen Transformation. *Zeitschrift für Dialektologie und Linguistik*, (Online First):1–27.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal ..., and Barret Zoph. 2024. Gpt-4 technical report.

J. Alan Pfeffer. 1975. *Grunddeutsch: Erarbeitung und Wertung dreier deutscher Korpora. Ein Bericht aus dem Institut for Basic German, Pittsburgh*. Narr, Tübingen.

Jeffrey Pheiff and Tillmann Pistor. 2023. Des reine Hochdeutsch bring ich net zamm. Lautliche Merkmale des Regionalakzents im Ostfränkischen. *Wiener Linguistische Gazette*, 94:35–95.

Jeffrey Pheiff, Tillmann Pistor, and Anna Wolańska. 2019. Zur Verwendung der Forschungsplattform Regionalsprache.de (REDE) bei der Vermittlung arealer Sprachvariation in den Bereichen Deutsch als Fremdsprache und Deutsch als Zweitsprache. *Linguistik Online*, 94(1):101–137.

Tillmann Pistor. 2022. *Universelle Intonationsmuster. Ein empirischer Nachweis konstanter prosodischer Strukturen in Regionalsprachen des Deutschen und darüber hinaus*. Steiner, Stuttgart.

Tillmann Pistor. 2025a. Phonetische Merkmale regionalsprachlicher Prosodie in Deutschland. *Zeitschrift für Dialektologie und Linguistik*, 92(3):332–373.

Tillmann Pistor. 2025b. So. Weischt? Zur Interaktion von Prosodie, syntagmatischer Position und lexikalischer Semantik von Diskursmarkern im Niederalemannischen. In Susanne Oberholzer and Noemi Adam-Graf, editors, *Dialekt in Gesellschaft und Schule. Variation und Wandel in Gebrauch und Wahrnehmung des Alemannischen*, pages 277–289. Steiner, Stuttgart.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Josephine Marie Rocholl. 2015. *Ostmitteldeutsch – eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum*. Olms, Hildesheim.

Jürgen Erich Schmidt. 2017. Vom traditionellen Dialekt zu den modernen deutschen Regionalsprachen. In Deutsche Akademie für Sprache und Dichtung and Union der deutschen Akademien der Wissenschaften, editors, *Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache*, pages 105–143. Stauffenburg, Tübingen.

Jürgen Erich Schmidt and Joachim Herrgen. 2011. *Sprachdynamik: eine Einführung in die moderne Regionalsprachenforschung*. Schmitt, Berlin.

Jürgen Erich Schmidt, Joachim Herrgen, Roland Kehrein, Alfred Lameli, and Hanna Fischer. 2020–. Regionalsprache.de. Forschungsplattform zu den modernen Regionalsprachen des Deutschen.

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, ..., and Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 10:353–402.

Marlon Siewert. 2023. Automatic Speaker Recognition and Dialectal Variation. Master's thesis, Marburg University.

Nadja Spina and Alfred Lameli. 2024. Regional variation in pre-boundary lengthening from a horizontal and vertical perspective: Evidence from German dialect- and standard-targeted speech. In *Proceedings of the Conference on Speech Prosody 2024*, pages 990–994.

Nadja Spina and Alfred Lameli. 2025. Slow or Long? Generational Variation of Pre-boundary Lengthening in Different Dialect Areas of German. In *Proceedings of the Third International Conference on Tone and Intonation*, pages 60–64.

Rico Stiel. 2020. *Phonemwandel im gesprochenen Standard. Dynamik des //-Phonems im Deutschen*. Ph.D. thesis, Marburg University.

Lars Vorberger. 2019. *Regionalsprache in Hessen. Eine Untersuchung zu Sprachvariation und Sprachwandel im mittleren und südlichen Hessen*. Steiner, Stuttgart.

Peter Wiesinger. 1983. Die Einteilung der deutschen Dialekte. In Werner Besch, Ulrich Knoop, Wolfgang Putschke, and Herbert Ernst Wiegand, editors, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, volume 1.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 807–900. De Gruyter, Berlin and New York.

Qin Yan and Saeed Vaseghi. 2002. A comparative analysis of uk and us english accents in recognition and synthesis. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–413–I–416.

Eberhard Zwirner. 1962. *Deutsches Spracharchiv 1932–1962. Geschichte, Aufgaben und Gliederung, Bibliographie*. Aschendorff, Münster (Westfalen).

A Catalog of Basque Dialectal Resources: Online Collections and Standard-to-Dialectal Adaptations

Jaione Bengoetxea, Itziar Gonzalez-Dios, Rodrigo Agerri
HiTZ Center - Ixa, University of the Basque Country EHU
{jaione.bengoetxea,itziar.gonzalezd,rodrigo.agerri}@ehu.eus

Abstract

Recent research on dialectal NLP has identified data scarcity as a primary limitation. To address this limitation, this paper presents a catalog of contemporary Basque dialectal data and resources, offering a systematic and comprehensive compilation of the dialectal data currently available in Basque. Two types of data sources have been distinguished: online data originally written in some dialect, and standard-to-dialect adapted data. The former includes all dialectal data that can be found online, such as news and radio sites, informal tweets, as well as online resources such as dictionaries, atlases, grammar rules, or videos. The latter consists of data that has been adapted from the standard variety to dialectal varieties, either manually or automatically. Regarding the manual adaptation, the test split of the XNLI Natural Language Inference dataset was manually adapted into three Basque dialects: Western, Central, and Navarrese-Lapurdian, yielding a high-quality parallel gold standard evaluation dataset. With respect to the automatic dialectal adaptation, the automatically adapted physical commonsense dataset (BasPhyCo$_{west}$) underwent additional manual evaluation by native speakers to assess its quality and determine whether it could serve as a viable substitute for full manual adaptation (i.e., silver data creation).

Keywords: Basque, dialects, low-resource, data-collection

1. Introduction

Dialectal variation is a core feature of all natural languages. However, up until now, Natural Language Processing (NLP) research has almost exclusively focused on tailoring data and resources for the standard forms of each language.

In recent years, this trend has slowly started to shift, with some studies increasingly focusing on dialects. However, the range of tasks and languages addressed remains fairly limited. Several surveys review recent developments and research directions in dialectal NLP, such as Joshi et al. (2025) and Zampieri et al. (2020).

Therefore, a major limitation identified by these works is the scarcity of data and resources for non-standard varieties. This presents a significant challenge, since recent advances in state-of-the-art NLP have reinforced the importance of data quantity in developing high-performing language technology tools, especially in low-resource scenarios (Artetxe et al., 2022). Given the importance of data quantity, the field of dialectal NLP could be regarded as a low-resource research scenario.

The lack of modern dialectal data is especially pronounced for Basque NLP technologies. The majority of resources have been developed with a high focus on Standard Basque, such as spell-checkers (Agirre et al., 1992), Neural Machine Translators[1] or, more recently, text representations (Agerri et al., 2020), and instruction fine-tuned Large Language Models (LLMs) trained for Basque (Etxaniz et al., 2024; Sainz et al., 2025).

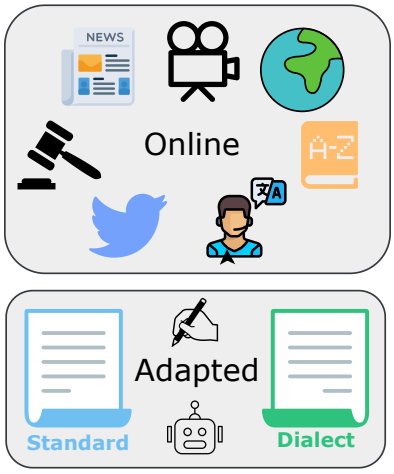


Figure 1: The Basque dialectal catalog consists of two different sources: **online** dialectal data and standard-to-dialect **adapted** data, either manually or automatically.

In this context, this paper aims to present a thorough compilation of Basque dialectal data and resources in order to facilitate potential future work on Basque dialects. We hope this will be useful not only in the field of NLP, but also in other related areas such as sociolinguistics or variationist linguistics.

This work groups Basque dialectal data and resources into two main categories, presented in Section 3 and Section 4, respectively:

[1] Elia or Itzuli.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 153–164
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

- **Online dialectal data and resources.** This category includes online sites such as local news and radio stations, tweets, linguistic atlases, dictionaries, and videos.
- **Standard-to-dialect adapted data.** This includes standard data adapted into dialects, either manually or automatically. For the former, we present **Parallel XNLIvar**, an expansion of XNLIvar (Bengoetxea et al., 2025) that includes the manual adaptation of the 5000-instance test set for the task of Natural Language Inference (NLI). The latter describes **BasPhyCo**, a physical commonsense dataset automatically adapted into dialectal Basque using Large Language Models (LLMS) (Bengoetxea et al., 2026). A novel manual evaluation has been conducted by native dialect speakers on the automatically obtained data.

These resources and adapted datasets are collected and publicly available[2].

2. Related Work

Dialectal Data Collection Many recent studies have focused their research on the systematic collection of dialectal data as a strategy to address the problem of data scarcity. For instance, Sun et al. (2025) introduced a gamified dialect data collection system, where native speakers could access an interface and either rewrite sentences into their dialect or match dialectal sentences to their geographical location. Their work has been found to efficiently increase user engagement in data collection processes.

However, given the growing demand to acquire large volumes of data, manual collection is frequently impractical. Therefore, several works have resorted to silver data creation, such as MultiVALUE (Ziems et al., 2023), a rule-based translation system for 50 different English dialects and 189 linguistic features, which performs as a successful data augmentation method.

Although the critical role of dialectal data is widely recognized, no study currently provides a comprehensive dialectal database. There are, however, some benchmark works that do not explicitly focus on data, but provide an invaluable source of dialectal data. For instance, Faisal et al. (2024) presented an extensive dialectal benchmark, evaluating 10 different tasks in 281 varieties. Additionally, Alam et al. (2024) focused their dialectal benchmark on Machine Translation (MT), as they evaluated several varieties from 12 different languages.

Both of these benchmarks included Basyque (Uria and Etxepare, 2012), a dataset of Northern Basque dialects used to evaluate MT. No other task was evaluated in Basque in these works, as there was no other dialectal dataset for modern Basque dialects at the time of these benchmarks.

Basque Dialects The classification of Basque dialects has been up for debate for many years. Bonaparte (1869) proposed a classification of eight dialects and 25 subdialects, a distribution that highlights the remarkable degree of variation within Basque, despite its comparatively small geographic area.

This classification has been considered canonical until a recent Basque dialectology work by Zuazu (2008), which established an extensive and comprehensive categorization of five Basque dialects, and provided a broad archive of the most representative features of each variation.

This paper will adopt the dialectal classification proposed by Zuazu (2008), distinguishing the following Basque dialects: Western, Central, Navarrese, Navarrese-Lapurdian and Zuberoan. For the purposes of this work, the Navarrese-Lapurdian and Zuberoan dialects will additionally be referred to as Northern dialects. A map illustrating this classification by Zuazu (2008) is provided in the Appendix.

Basque Standardization While extensive dialectal diversity may constitute a source of linguistic richness, it can also increase the risk of language endangerment. This concern led many Basque linguists to push for the necessity of a standard variation (Garabide, 2010).

The process for the creation of a standard variation was a lengthy one. Many meetings were held by contemporary linguists, until Koldo Mitxelena's proposal in the 60s. He suggested using the Central dialects as the foundation of the Standard, mainly due to its practicality: it was the dialect with the most literary prestige to date, as well as being understandable by all Basque speakers. This proposal was discussed, modified, and accepted in the Congress of Arantzazu in 1968 (Garabide, 2010).

The emergence of the standard variety was followed by growing movements to preserve and revitalize dialects. For instance, the Standard Western variety, which was developed to formalize and thus promote its use in written form as well as other registers and use cases (Labayru and Kutxafundazioa, 2001).

Overall, Basque standardization and dialect preservation have developed side by side, both playing an important role in shaping the language today. This work aims to contribute to this movement by providing a collection of current dialectal resources.

[2] https://github.com/hitz-zentroa/Catalog-of-Basque-Dialects

Source	Type	Dialect	Register	Modality	License
Bizkaia Irratia	Text & Audio	Western	Formal	News & Radio	cc-by-sa
Bizkaie!	Text & Audio	Western	Formal	News & Radio	CC-BY-SA
Xiberoko Botza	Text & Audio	Zuberoan	Formal	News & Radio	CC-BY-SA
Irulegiko Irratia	Text & Audio	Nav-Lap	Formal	News & Radio	CC-BY-SA
Gure Irratia	Text & Audio	Nav-Lap	Formal	News & Radio	CC-BY-SA
General Assemblies	Text	Western	Specialized	Minutes	N/A
BSM	Text	Mixed	Informal	Tweets	CC-BY-SA
Linguistic Atlas	Resource	Mixed	-	Atlas	N/A
Euskaltzaindia	Resource	Mixed	-	Dict.	N/A
LabayruHiztegia	Resource	Western	-	Dict.	N/A
LabayruGramatika	Resource	Western	-	Grammar	N/A
Basyque	Resource	Northern	-	Grammar	CC-BY-SA
Ahotsak	Video	Mixed	-	Speech	CC-BY-SA
Mintzoak	Video	Northern	-	Speech	CC-BY-NC-SA
Euskalkiak.eus	Video	Mixed	-	Speech	CC-BY-SA

Table 1: Summary of online Basque dialectal resources. N/A = Not Available.

3. Online Dialectal Data

In this section, we introduce the dialectal data available within the Basque online community, categorized according to their modality. All sources and their characteristics are summarized in Table 1.

3.1. News and Radio Sites

Several Basque news sites write articles in their local dialectal variety. For instance, news sites that have been written in the Western dialect include **Bizkaia Irratia** and **Bizkaie!**. Inside the news domain, these texts are written in a formal, not technical register.

These sites provide up-to-date news articles in the Standard Western dialect. This variety was first formalized in Labayru and Kutxafundazioa (2001), where orthography, morphology (including verb declination), and syntax issues were established. More recently, the Labayru foundation has digitalized this grammar and provided Standard Western Basque information on their website (more information on Section 3.6).

Apart from the Western dialect, several online news sites are also available in other varieties, including the Northern dialects. In fact, the **Euskal Irratiak** association consists of several independent radios that collect news from different locations, in their local variation of Basque. These include **Xiberoko Botza**, **Irulegiko Irratia** and **Gure Irratia**.

3.2. Legal Documents

The website of the **Biscayan General Assemblies** contains the minutes of the highest organizational body representing the citizens of Biscay, i.e., the Biscayan parliament. This organization exercises regulatory power in the region and approves the budgets for the historical territory.

Some of the legal documents on their website are written in the Western dialect, providing dialectal texts not only in a formal register, but also in the specialized legal domain. As is the case for the news site texts, these technical documents are written in the Standard Western dialect.

3.3. Basque Social Media Corpus

Some previous work has been done on the comparison of different registers in Basque, where informal speech often includes strong dialectal variations. These works include Fernandez de Landa et al. (2019) and Fernandez de Landa and Agerri (2021), where they work on real-world data collected from Twitter, which includes dialectal, slang, informal, and code-switched data. The **Basque Social Media (BSM)** corpus consists of approximately 11 million posts produced by more than 13,000 Basque-speaking users, with a total of around 188 million words.

3.4. Linguistic Atlas

Euskaltzaindia is the academic regulatory institution for the Basque language. In their many efforts to research and support the language, they have conducted some invaluable work on dialects, including the **Linguistic Atlas of Basque**.

This atlas provides linguistic maps for many Basque words according to their geographical location, thus illustrating the big lexical variation that could be encountered through the several Basque-

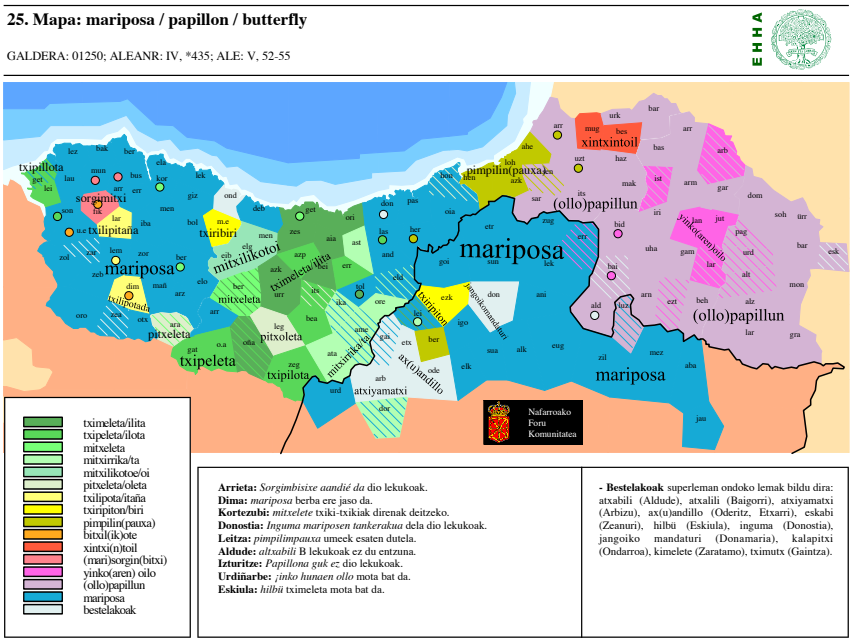


Figure 2: Example of a linguistic atlas.

speaking regions. An example map is provided in Figure 2.

3.5. Dictionaries

Some online dictionaries often include dialectal information when searching for specific words, such as Elhuyar or **Euskaltzaindia**. The Elhuyar dictionary provides dialectal information about some words. The Euskaltzaindia dictionary provides not only geographical information for selected words, indicating their potential dialectal affiliation, but also enables searches to be filtered by dialect.

However, these standard dictionaries only provide information on dialectal words that are accepted in the standard, while other variations of dialectal words are absent. For instance, *berba* (word) is accepted in the standard form and appears in dictionaries, while its possible variations, such as *berbie* or *berbi*, do not.

Additionally, some dictionaries are designed to focus on a particular dialect, for example **LabayruHiztegia**. This is a Western-focused dictionary, with Standard-Western word pairs, as well as information on some multi-word expressions particular to this dialect.

3.6. Grammar

Some efforts have been made to formalize the grammar rules from several Basque dialects. For example, the Labayru foundation has recently launched **LabayruGramatika**, a compilation of the grammatical rules from the Western dialect, with explanations as well as examples for every grammatical phenomenon.

Furthermore, Uria and Etxepare (2012) provided **Basyque**, an online resource to store, organize, manage and search for all the information concerning dialectal variation in, specifically, the North-Eastern Basque dialects, providing information that enables the syntactic analysis of these dialects.

3.7. Dialectal Videos

In an attempt to observe, preserve and analyze diachronic variation, some speakers have been interviewed, and the recordings have been uploaded to several web pages. These videos are annotated with the speakers' geographical origin.

For instance, **Ahotsak** provides thousands of videos of interviews with people from different generations, genders, backgrounds, and geographical locations. Some of these interviews have been transcribed and could act as a great resource for the analysis of oral Basque dialects. However, the videos with missing transcriptions still pose a great challenge.

Similar to Ahotsak, **Mintzoak** compiles video interviews of Northern Basque speakers, in order to keep the collective memory alive. Contrary to Ahotsak, this site does not contain transcriptions of the interviews.

Additionally, **Euskalkiak.eus** also contains some videos based on the geographical location of the speakers, also with no transcriptions.

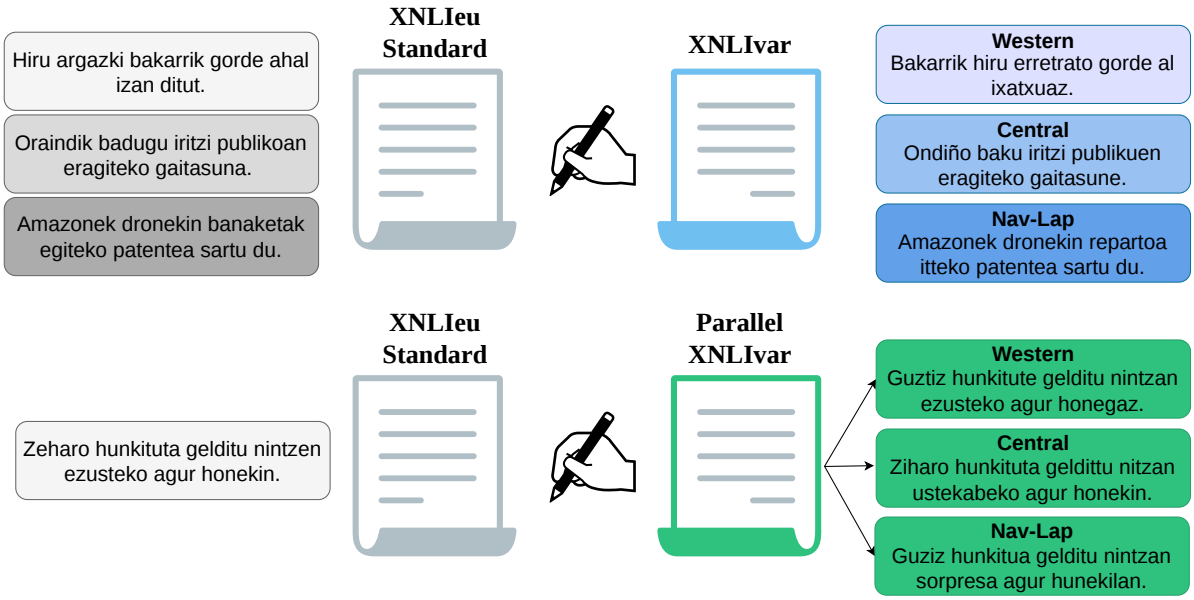


Figure 3: Illustration of XNLIeu dialectal adaptation. XNLIvar was a compilation of different instances in three different dialects. Parallel XNLIvar provides the same instances in three Basque dialects, offering completely parallel data.

4. Standard-to-Dialect Adapted Data

In addition to these online dialectal resources, both manual and automatic adaptations from standard Basque into dialects have recently been developed. Standard-dialect parallel data constitute an important resource not only for NLP but also for studies in variationist linguistics and sociolinguistics.

4.1. Manual Dialectal Adaptation

Manually adapting standard texts into dialects is to this day the most effective way to obtain gold standard data. However, despite its importance, relatively few studies have addressed this topic in Basque. This section describes the first manually adapted dataset into Basque dialects (Bengoetxea et al., 2025). Additionally, this paper presents a newly developed expansion of it, namely Parallel XNLIvar, also collected through manual adaptation.

XNLIvar Bengoetxea et al. (2025) presented XNLIvar, a Natural Language Inference (NLI) dataset which contained data in three Basque dialects. This dataset was manually adapted from XNLIeu (Heredia et al., 2024), an NLI dataset in Basque translated from the multilingual XNLI dataset (Conneau et al., 2018). This dataset consists of Premise-Hypothesis pairs with entailment, neutral or contradiction relations.

For the dialectal adaptation, Bengoetxea et al. (2025) adapted the native partition of XNLIeu, i.e., a 621-instance test set manually created in Basque. Although the dataset contained material from three Basque dialects (Western, Central, and Navarrese), it did not include fully parallel versions of the same content across dialects.

Dialect	Birthplace	Age	Gender	Studies
Western	Elorrio	58	Male	Translation
Central	Arroa/Zumaia	34	Female	Translation
Nav-Lap	Donibane-Lohizune	33	Female	Translation

Table 2: Metadata from native Basque speakers who adapted the test partition of XNLIeu into dialects.

Parallel XNLIvar This paper presents an expansion of XNLIvar by adapting the original XNLI test set of around 5000 instances into three different Basque dialects (Western, Central, and Navarrese-Lapurdian). This expansion was created in parallel, providing three fully dialectal versions of the XNLIeu test set, each corresponding to a different Basque dialect. This adaptation process is illustrated in Figure 3.

The adaptation from standard to dialectal Basque has been manually carried out by native speakers of each dialect. Metadata from each native speaker is illustrated in Table 2, such as birthplace, age, gender and previous studies.

All in all, this novel resource allows for a more thorough assessment of dialectal effects in NLI and facilitates per-dialect analysis through the inclusion of parallel data.

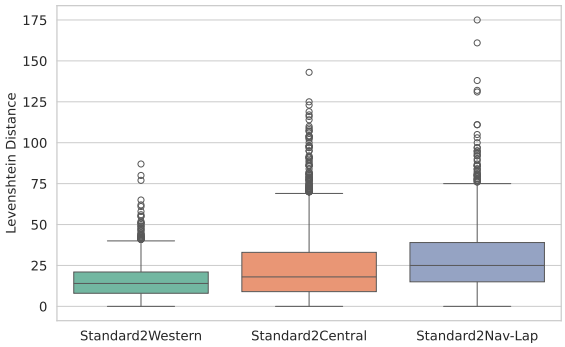


Figure 4: Levenshtein distance distribution for the three Parallel XNLIvar datasets.

4.1.1. Analysis of Parallel XNLIvar

Further analysis has been conducted to explore how different the dialectal datasets are from the standard. To do so, the Levenshtein distance between standard and dialectal sentences was calculated. This metric measures the minimum number of single-character insertions, deletions, or substitutions required to transform one string into another. The average distance results are presented in Table 3.

The results suggest that Navarrese-Lapurdian is the dialect most distant from the standard variety. This is consistent with Basque dialectological research, which has shown that peripheral dialects typically exhibit greater divergence (Mitxelena, 1981).

Under this theoretical framework, the Western dialect should not be the closest to the standard variety, given that it is likewise considered a peripheral dialect. However, as we can see in Figure 4, the Central dialect seems to be further from the standard than the Western. This could be due to the distance distribution, as the Central dialect shows many sentences with big modifications, whereas the distances in the Western dataset remain consistently low across all sentences.

Further analysis of the two datasets shows that the linguist who did the Central adaptations frequently modified sentence word order. That could be why the Levenshtein distance for the Central dataset is higher than for the Western dataset, in which word order was largely preserved. This highlights the importance of robust metrics for measuring dialectal variation, as this distance metric estimates the transformation distance between sentences, but its effectiveness in capturing dialectalness seems to remain unclear, especially in free word-order languages like Basque.

4.2. Automatic Dialectal Adaptation

Given the high cost and time demands of manual dialectal adaptation, recent work has focused on automatically converting standard data into dialects, which is discussed in this section.

BasPhyCo Bengoetxea et al. (2026) presented a physical commonsense reasoning dataset consisting of 356 instances of 5-sentence stories, which could be plausible or implausible. BasPhyCo was manually translated into standard Basque from its original Italian version (Pensa et al., 2024). Additionally, the standard Basque dataset was automatically adapted into the Western dialect through few-shot prompting of Latxa-It-70B (Sainz et al., 2025).

Consequently, Bengoetxea et al. (2026) provide two parallel versions of the same dataset: BasPhyCo and BasPhyCo$_{\text{west}}$. The availability of standard–dialect parallel data allows to examine the impact of dialectal variation on physical commonsense reasoning.

4.2.1. Evaluation of BasPhyCo$_{\text{west}}$

The automatic adaptation of BasPhyCo$_{\text{west}}$ was validated by a professional Basque linguist (Bengoetxea et al., 2026). However, we sought to extend this evaluation to the adapted dataset, thereby measuring the actual dialectal value of the automatic adaptation for native Western dialect speakers. To do so, we have outlined the following manual evaluation framework.

Evaluated datasets The original prompt of Bengoetxea et al. (2026) for the adaptation of BasPhyCo$_{\text{west}}$ explicitly allowed for non-standard orthographic modifications. We have adapted this prompt by eliminating the possibility of orthographic changes and obtained a new version, BasPhyCo$_{\text{west-new}}$. Both prompts used for the adaptations are provided in the Appendix.

Therefore, our evaluation was done on three versions of the dataset: the **Standard** (BasPhyCo), the Western **Orthographic** (BasPhyCo$_{\text{west}}$), and a novel **Western**, non-standard orthographic version (BasPhyCo$_{\text{west-new}}$). The following examples illustrate two dialectal sentences, both with standard and non-standard orthography.

Dialect	Distance
Western	15.56
Central	24.73
Nav-Lap	29.36

Table 3: The average Levenshtein **Distance** of the Parallel XNLIvar dialectal datasets to the Standard.

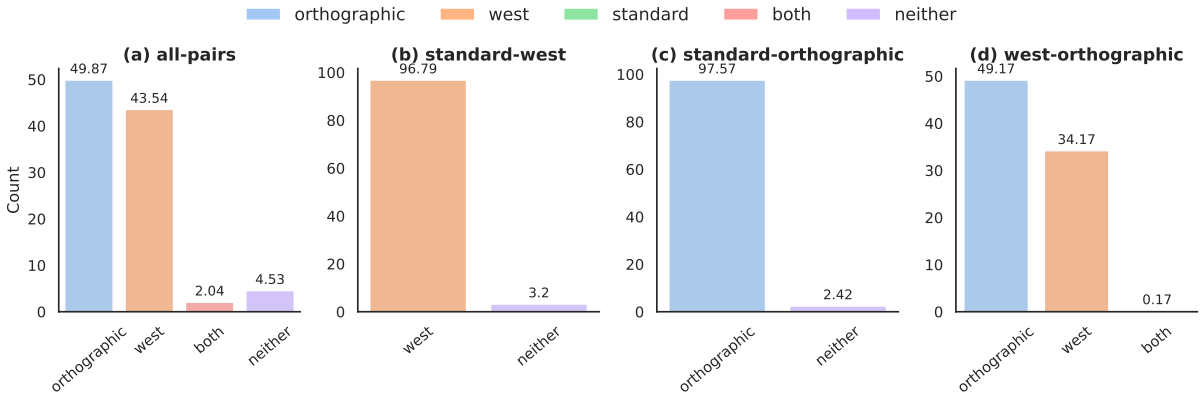


Figure 5: Manual evaluation results. From left to right, results for all sentence pairs (**all-pairs**), as well as results for different sentence pair combinations (**standard-west**, **standard-orthographic**, and **west-orthographic**).

Standard Teknikaria ez da oraindino etorri

Non-Standard Teknikarixa ez dau oraindiño etorri

We can observe that the non-standard version includes the letter *ñ* in *oraindiño*, while the standard version follows standard orthographic rules and favors the use of the letter *n*.

Description of the task The evaluation was proposed as a pairwise comparison task. The pairs of sentences to be evaluated were constructed as a Cartesian product, i.e., every sentence from dataset A was paired with every sentence in dataset B. Thus, the combination of datasets A $\times$ B is the following:

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}. \quad (1)$$

As we have three different datasets (A, B and C), the Cartesian product for all three dataset combinations was conducted and subsequently concatenated. Given that $\times$ is the Cartesian product, the pairs to be evaluated were constructed as:

$$(A \times B) \cup (A \times C) \cup (B \times C) \quad (2)$$

For the manual evaluation, 180 sentence pairs from each dataset combination were randomly sampled, obtaining a total of 540 sentence pairs for evaluation.

The task consisted of annotating which sentence from each pair was closer to the Western dialect of Basque. The evaluators were given four options: *sentence A* is more Western, *sentence B* is more Western, they are *both* Western, or *neither* is Western.

Dialect	Birthplace	Age	Gender	Studies
Western	Oñati	33	Female	Translation
Western	Elorrio	59	Male	Translation

Table 4: Metadata from the native Western Basque speakers for the manual evaluation of the automatic adaptation.

Annotators The evaluation was made by two different native Western speakers, with the same educational background, but different age, gender, and birthplace. The metadata for each evaluator is presented in Table 4.

The Inter-Annotator Agreement (IAA) between the two evaluators has been computed, with a Cohen's Kappa of 0.71, which constitutes a substantial agreement score.

Quantitative results The general results, as well as the results for different sentence-pair types, are illustrated in Figure 5.

The results for all sentence types (5a) show that the orthographic dataset has the most Western features, although it is closely followed by the Western dataset. No Standard sentence was marked as dialectal, which consolidates the confidence of the evaluation quality. Evaluators marked some doubtful sentence pairs as *Neither* or *Both*, which are examined in the following analysis section.

Regarding the results for the different sentence-pair types, we can observe that when the dialectal sentences were paired with a standard sentence (i.e., standard-west and standard-orthographic sentence pairs), the evaluators always chose the adapted sentences over the standard (Figure 5b and 5c). Additionally, when having to choose between the two adapted sentence types (west-orthographic), the evaluators deemed the Orthographic dataset more dialectal, but still closely fol-

lowed by Western sentences (5d).

Consequently, this manual evaluation has shown that both automatically adapted datasets (Orthographic and Western) contain considerable dialectal features compared to the standard version. Additionally, explicitly stating in the adaptation prompt that non-standard orthographic changes are possible seems to generate even more dialectal adaptations, according to this evaluation.

Analysis During the pairwise comparison, evaluators were also given two extra options: *Both* and *Neither*. It can be observed in Figure 5b and 5c that all *Neither* labeled pairs seem to occur when one of the sentences in the pair is from the standard dataset. This reveals that the automatic adaptation sometimes failed to transform the standard sentences into their dialectal form. For example, the following sentences from the adapted datasets do not contain dialectal features:

Ortho Koldo esnatu da. (*Koldo has woken up*)

West Izotz-ontziak urtu dira. (*The ice-cubes have melted*)

Similarly, all *Both* instances seem to occur when comparing the two automatically adapted datasets. This highlights that the Standard dataset is not biased towards dialectal language, as evaluators have not once considered a standard sentence to be dialectal.

Furthermore, disagreements between annotators have also been examined. The majority of these instances occur on difficult sentence pairs, i.e., sentences that contain little to no dialectal features. Further analysis of these instances has revealed a slight annotator bias: one annotator considered *heldu* (arrive) a Western marker, whereas the other did not.

Ortho Ane berandu **helduko** da etxera. (*Ane arrived late home*)

Standard Ane berandu **iritsiko** da etxera. (*Ane arrived late home*)

Finally, although error identification was not an annotation requirement, evaluators noted several errors in the dialectal adaptations. This points to clear limitations in the current dialectal adaptation capabilities of LLMs, which require further investigation.

5. Conclusion

This work presents a comprehensive collection of Basque dialectal data, categorized into two groups. First, online dialectal data and resources have been presented, grouped according to their domain, such as news, legal documents, informal tweets, dictionaries, grammar collections or even audiovisual resources. Secondly, standard datasets that were adapted into dialects have been described, both manually and automatically.

This paper has introduced two main contributions: (i) Parallel XNLIvar, a novel manually adapted parallel dialectal dataset for NLI. (ii) A manual evaluation of the automatically adapted $\text{BasPhyCo}_{\text{west}}$ dataset, outlined as a pairwise comparison task and performed by native dialectal speakers.

Limitations

The main limitation regarding online sources is that some data cannot be used due to licensing issues.

Additionally, automatic dialectal adaptation has been evaluated for only one Basque dialect, while other dialects remain unexamined.

Finally, the manual evaluation of the automatic dialectal adaptation has revealed some errors, which present an evident limitation in the possibility of silver data creation through automatic adaptation.

Acknowledgments

This work has been supported by the HiTZ center and the Basque Government (Research group funding IT-1805-22). Jaione Bengoetxea is funded by the Basque Government pre-doctoral grant (PRE_2024_1_0028).

We also acknowledge the following MCIN/AEI/10.13039/501100011033 project: (i) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR and (ii) DeepThought (PID2024-159202OB-C21) funded by ERDF, EU.

Bibliographical References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788.

Eneko Agirre, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Montserrat Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on two-level morphology. In *Third Conference on Applied Natural Language Processing*, pages 119–125, Trento, Italy. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2025. Lost in variation? evaluating NLI performance in Basque and Spanish geographical variants. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 452–468, Vienna, Austria. Association for Computational Linguistics.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2026. Physical commonsense reasoning for lower-resourced languages and dialects: a study on basque.

Louis-Lucien Bonaparte. 1869. Le verbe basque en tableaux. *Berrargitaratua:[JA Arana Martija, arg.], Opera omnia vasconice*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Fahin Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialect-bench: A nlp benchmark for dialects, varieties, and closely-related languages. *ArXiv*, abs/2403.11009.

Joseba Fernandez de Landa and Rodrigo Agerri. 2021. Social analysis of young basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0):1–15.

Joseba Fernandez de Landa, Rodrigo Agerri, and Iñaki Alegria. 2019. Large scale linguistic processing of tweets to understand social interactions among speakers of less resourced languages: The basque case. *Information*, 10(6).

Elkartea Garabide. 2010. Hizkuntzaren estandarizazioa. *Euskararen berreskuratzea ii*.

Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Jeremy Barnes, and Aitor Soroa. 2024. XNLIeu: a dataset for cross-lingual NLI in Basque. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4177–4188, Mexico City, Mexico. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

Ikastegia Labayru and Bilbao Bizkaia Kutxafundazioa. 2001. Bizkai euskeraren jarraibide liburua. *Lehenengo Pausuak. Labayru Ikastegia-BBK-Eusko Jaurlaritza-Bizkaiko Foru Aldundia-Bizkaia Irratia*.

Luis Mitxelena. 1981. Lengua común y dialectos vascos. *Anuario del Seminario de Filología Vasca" Julio de Urquijo"*, 15:289–313.

Giulia Pensa, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. A Multi-layered Approach to Physical Commonsense Understanding: Creation and Evaluation of an Italian Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 819–831.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. Instructing large language models for low-resource languages: A systematic study for Basque. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29136–29160, Suzhou, China. Association for Computational Linguistics.

Jiugeng Sun, Rita Sevastjanova, Sina Ahmadi, Rico Sennrich, and Mennatallah El-Assady. 2025. Dia-lingle: A gamified interface for dialectal data

collection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 148–158, Vienna, Austria. Association for Computational Linguistics.

Larraitz Uria and Ricardo Etxepare. 2012. Hizkeren arteko aldakortasun sintaktikoa aztertzeko metodologiaren nondik norakoak: Basyque aplikazioa. *Lapurdum. Euskal ikerketen aldizkaria| Revue d'études basques| Revista de estudios vascos| Basque studies review*, (16):117–135.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

Koldo Zuazu. 2008. *Euskalkiak. Euskararen dialektoak*. Elkar.

A. A Map of Basque Dialects

The map in Figure 6 shows the Basque dialect classification according to Zuazu (2008).

B. Dialectal Adaptation Prompts

The prompts used to prompt Latxa-3.1-70B-It for the dialectal adaptation are provided below, both for the non-standard orthographic changes (Figure 7) and for the standard orthographic changes (Figure 8).



Figure 6: The map of the classification of Basque dialects according to Zuazu (2008).

I will give you three versions of a story. Each version has five sentences. Some sentences are identical across versions. You need to adapt this text so that it includes Bizkaian dialectal features. **You can use non-standard orthography. Try to make it as similar as possible to oral language.**

Task:

1. First, list all unique sentences across all three stories.
2. Adapt each unique sentence exactly once into the Bizkaian dialect.
3. Then reconstruct the three stories with the translations, making sure that any identical source sentence always has the identical translation.
4. If there are more than three stories, repeat the same process for all of them.

Format:

This is an example of an standard (INPUT) instance and an example of the dialectal (OUTPUT) adaptation that you need to do:

Standard:

STORY1: ['Mikel lanera joan da', 'Mikelek ordenagailua piztu du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']
STORY2: ['Mikel lanera joan da', 'Mikelek mezuak erantzun ditu', 'Mikelek mezuak irakurri ditu', 'Mikelek ordenagailua piztu du', 'Mikel etxera joan da']
STORY3: ['Mikel lanera joan da', 'Mikelek ordenagailua itzali du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']

Dialectal:

STORY1: ['Mikel lanera jun de', 'Mikelek ordenagaillua piztu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']
STORY2: ['Mikel lanera jun de', 'Mikelek mesuek erantzun ditu', 'Mikelek mesuek irakurri dauz', 'Mikelek ordenagaillua piztu dau', 'Mikel etxera jun de']
STORY3: ['Mikel lanera jun de', 'Mikelek ordenagaillua amatatu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']

Output only the reconstructed stories in the exact same format as the input. Do not output explanations, steps, or commentary.

Figure 7: Dialectal adaptation prompt for the non-orthographic dataset version ($\text{BasPhyCo}_{\text{west}}$).

I will give you three versions of a story. Each version has five sentences. Some sentences are identical across versions. **You need to adapt this text so that it includes Bizkaian dialectal features.**

Task:

1. First, list all unique sentences across all three stories.
2. Adapt each unique sentence exactly once into the Bizkaian dialect.
3. Then reconstruct the three stories with the translations, making sure that any identical source sentence always has the identical translation.
4. If there are more than three stories, repeat the same process for all of them.

Format:

This is an example of an standard (INPUT) instance and an example of the dialectal (OUTPUT) adaptation that you need to do:

Standard:

STORY1: ['Mikel lanera joan da', 'Mikelek ordenagailua piztu du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']
STORY2: ['Mikel lanera joan da', 'Mikelek mezuak erantzun ditu', 'Mikelek mezuak irakurri ditu', 'Mikelek ordenagailua piztu du', 'Mikel etxera joan da']
STORY3: ['Mikel lanera joan da', 'Mikelek ordenagailua itzali du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']

Dialectal:

STORY1: ['Mikel lanera jun de', 'Mikelek ordenagaillua piztu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']
STORY2: ['Mikel lanera jun de', 'Mikelek mesuek erantzun ditu', 'Mikelek mesuek irakurri dauz', 'Mikelek ordenagaillua piztu dau', 'Mikel etxera jun de']
STORY3: ['Mikel lanera jun de', 'Mikelek ordenagaillua amatatu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']

Output only the reconstructed stories in the exact same format as the input. Do not output explanations, steps, or commentary.

Figure 8: Dialectal adaptation prompt for the orthographic dataset version ($\text{BasPhyCo}_{\text{west-new}}$).

WoVis: Interactive Visualization of Word Embeddings for Semantic Change in Historical and Dialectal Language Resources

Filip Miletić[1] Maximilian Henkel[1,2] Rene Cutura[2] Sophie Sadler[2]
Quynh Quang Ngo[2] Michael Sedlmair[2] Sabine Schulte im Walde[1]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]VISUS, University of Stuttgart, Germany
[1]{filip.miletic, maximilian.henkel, schulte}@ims.uni-stuttgart.de
[2]{rene.cutura, sophie.sadler, quynh.ngo, michael.sedlmair}@visus.uni-stuttgart.de

Abstract

Computational modeling of language variation and change often relies on comparisons of word embeddings induced from existing historical and dialectal language resources. However, their use in the wider linguistics research community and in application domains such as lexicography is challenged by their limited manipulability for non-technical users, which in turn exacerbates the underuse of such resources. Aiming to foster a broader uptake of embedding-based analyses, we introduce WoVis, an interactive visualization tool designed to compare word embedding models in analyses of semantic change. Our system supports simultaneous model comparisons along two dimensions (e.g., language varieties and time periods) and provides analyses at different levels of granularity: an overview of the full vocabulary across all word embedding models, distributional behavior of individual words, targeted comparisons of word pairs, and model-external lexical features such as frequency and affective norms. We illustrate the utility of our system on two languages, German and English, with analyses of word usage across language varieties as well as time: West vs. East Germany, 1950–1989; and general-domain US vs. scientific UK English, ca. 1800–2000.

Keywords: word embeddings, semantic change, visualization, dimensionality reduction

1. Introduction

Computational modeling of language variation and change often relies on comparisons of word embeddings, i.e., high-dimensional representations of word meaning which reflect cooccurrence patterns in large corpora. Word embeddings are often induced from existing historical and dialectal language resources, and have been used to analyze phenomena including non-standard lexical choices in minority linguistic communities (e.g., Shoemark et al., 2018; Del Tredici et al., 2019; Miletić et al., 2021; Würschinger and McGillivray, 2025); distinctive properties of specialized communication as opposed to the general-domain standard (e.g., Schlechtweg et al., 2019; Bizzoni et al., 2020; McGillivray et al., 2025); and broad trends of semantic change over long periods of time (e.g., Sagi et al., 2009; Gulordava and Baroni, 2011; Hamilton et al., 2016; Schlechtweg et al., 2020; Maurer et al., 2023; Mahdizadeh Sani et al., 2024). But outside of computational research – for example, in linguistics at large and in downstream application domains such as lexicography – the use of embedding-based methods is challenged by their limited manipulability for non-technical users. This issue in turn exacerbates the underuse of existing historical and dialectal language resources, a more general problem particularly affecting lower-resourced languages and language varieties (Soria et al., 2013).

A broader uptake of embedding-based analyses can be supported with visualization systems (Jatowt et al., 2021). This strategy is already incipient in the two-dimensional projections often used to illustrate proposed word embedding analyses, where a target word's trajectory is usually depicted with respect to its nearest neighbors from different models (e.g., Hamilton et al., 2016; Del Tredici and Fernández, 2017; Yao et al., 2018; Rassem et al., 2024; McGillivray et al., 2025). But since these visualizations aim to showcase specific analyses in their language of interest, they are not designed for deployment with other language resources or for interactive exploration – vital factors for novel linguistic insights – nor do they extensively consider the choice of dimensionality reduction techniques. Some of these issues are overcome by general-purpose embedding visualization tools with an interactive interface and multiple dimensionality reductions (e.g., TensorFlow Embedding Projector; Smilkov et al., 2016), but these are not designed for linguistic research and, in particular, lack comparisons of multiple models at the same time. Both analysis-specific and general-purpose visualizations present another important shortcoming – the limited integration of model-external lexical features such as word frequencies and affective properties. These have been shown to correlate with language change (Blank, 1999; Bybee, 2015; Jenkins, 2026) and can therefore aid the interpretation of visualized trends.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 165–176
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

In order to address these challenges, we provide a series of contributions aiming to improve the accessibility of existing historical and dialectal language resources. **(1) We release WoVis, an interactive visualization tool designed to compare word embedding models** in analyses of semantic change (Figure 1). It supports simultaneous model comparisons along two dimensions (e.g., language varieties *and* time periods) and relies on PaCMAP (Wang et al., 2021), a dimensionality reduction technique which aims to preserve both the local and the global structure of the vector space. The system can be deployed on available word2vec models and also provides a module to train new models from available corpora. **(2) We identify and implement a diverse set of computational semantic analyses** which underpin WoVis, bringing together complementary perspectives on the semantic properties of individual words as well as their relatedness to the rest of the vocabulary. When our system is deployed, it first provides an overview of the full vocabulary across all word embedding models, and then analyzes the evolution of individual words in terms of distributional behavior (nearest neighbors and neighborhood densities), targeted comparisons with other words (capturing shifts in dominant senses), and model-external lexical features (frequency and affective norms). **(3) We show the utility of WoVis with sample analyses of existing resources in two languages**, German and English. We explore a challenging scenario which compares word usage across language varieties as well as time: West vs. East Germany, 1950–1990; and scientific UK vs. general-domain US English, ca. 1800–2000. Importantly, WoVis can also be applied out-of-the-box to compare two word embeddings models from other languages and datasets.[1]

The remainder of this paper first introduces the data (§2) and models (§3) used in our visualizations. We then present the implementation of our system (§4) and the insights from sample analyses (§5). Finally, we close with concluding remarks (§6).

2. Data

This section introduces the data resources underlying our visualizations. While the use of our system is not limited to any specific language or dataset, we showcase its performance on German and English. Both languages are equipped with numerous existing resources, which however may be underused due to limited practical accessibility. We illustrate how our system can help address this challenge, with a particular focus on better utilizing available historical and dialectal resources.

[1]Our code and models are available here: https://github.com/maxeljaxel/wovis

Spiegel		**Berliner Zeitung**	
1950–1959	107.6	1950–1959	400.5
1960–1969	180.8	1960–1969	555.3
1970–1979	228.6	1970–1979	580.2
1980–1989	289.6	1980–1989	618.6
RSC		**CCOHA**	
1800–1849	9.1	1810–1849	44.4
1850–1899	37.0	1850–1899	107.8
1900–1949	65.4	1900–1949	138.6
1950–1996	172.0	1950–2009	181.4

Table 1: Structure of German and English corpora with time periods and size in millions of tokens

For each language, we select two sample corpora which correspond to distinct language varieties and which record their evolution over a parallel time period. Each variety-specific corpus is partitioned into time slices; we refer to these variety–time partitions as *subcorpora*. Their structure is summarized in Table 1 and further described below. Beyond the subcorpora – which are subsequently used to train word embedding models – we further use model-external lexical features to enrich our analysis of target words, also presented below.

2.1. German corpora

Our German-language corpora target the period between 1949 and 1990, during which Germany was formally divided into two states: the Federal Republic of Germany (FRG; West Germany) and the German Democratic Republic (GDR; East Germany). While German was the only official language of both states, substantial political, economic, and social differences led to gradual divergence in language use between the two state-level varieties, especially in the lexicon (Kinne and Strube-Edelmann, 1981; Ammon, 1991; Hellmann, 1993). We aim to capture these divergence phenomena – which for several decades underpinned two competing linguistic standards – by comparing newspaper texts from West and East Germany. For both corpora, we use texts published from 1950 to 1989 and split them into four subcorpora.

Spiegel. As our West German corpus, we use texts from the Spiegel, a weekly news magazine founded in 1947 in Hanover and based in Hamburg since 1952. The magazine mostly covers political and societal topics, with a historical focus on West Germany as well as international events. Our corpus was compiled by crawling the entire digital archive of the Spiegel with permission.[2]

[2]We are grateful to the Spiegel for granting us access to their archive for the purposes of this research: https://www.spiegel.de/spiegel/print/

Berliner Zeitung. As our East German corpus, we use texts from the Berliner Zeitung, a daily newspaper founded in 1945 in East Berlin and published to date. From 1953 to 1990, it was controlled by the Socialist Unity Party of Germany, i.e., the ruling party in East Germany. As a daily newspaper, it has a stronger focus on local events than the Spiegel. We rely on the corpus digitized and published by Berlin State Library.[3]

2.2. English corpora

In our English-language analysis, we compare corpora capturing different national varieties as well as domains (scientific UK English vs. general-domain US English), and we examine them over a longer time span (≈1800–2000). This setup targets the evolution of scientific writing (Biber and Gray, 2016; Teich et al., 2016; Zaberer, 2025) in contrast to the general-domain standard. It is also less restricted than the German-language analysis in terms of time and domains, and as such constitutes a more challenging scenario for our visualization system.

RSC. As a corpus of UK English from the scientific domain, we use the Royal Society Corpus (RSC) (Kermes et al., 2016; Fischer et al., 2020). It contains texts published by the Royal Society of London from 1665 to 1996 in its Philosophical Transactions and Proceedings. We use the texts from 1800 to 1996, split into four subcorpora.

CCOHA. As a corpus of general-domain US English, we use the clean version of the Corpus of Historical American English (CCOHA) (Davies, 2012; Alatrash et al., 2020). It contains texts published between 1810 and 2009, which are moreover balanced across four genres (fiction, non-fiction, magazine articles, and news). We use all available data and once again split it into four subcorpora.

2.3. Model-external lexical features

In order to better explain the semantic change trends we observe, we additionally feed our visualization system with model-external lexical features which characterize individual words and are expected to correlate with semantic change (Blank, 1999; Bybee, 2015). We specifically include word frequency, which we calculate for each subcorpus; and psycholinguistic norms, taken from external resources. For German, we use abstractness, imageability, valence, and arousal norms (Köper and Schulte im Walde, 2016); for English, we use abstractness (Brysbaert et al., 2014) as well as valence, arousal, and dominance norms (Mohammad, 2018). Our visualization system is modular with respect to this type of information, i.e., it can accommodate different or additional lexical features relevant for the visualization task at hand.

[3] https://zefys.staatsbibliothek-berlin.de/ddr-presse/

3. Models

We now turn to the word embedding models used in our visualizations. We discuss the training and alignment procedure; the reduction into two dimensions for visualization; and the computation of neighborhood densities, which we later use to characterize word-level trajectories of semantic change.

3.1. Word embedding models

Following standard practice in computational research on semantic change, we use separate word2vec models (Mikolov et al., 2013) for each subcorpus. We deliberately opt for type-level rather than token-level word embeddings because they are far more compute-efficient and continue to provide robust results in linguistically oriented modeling (recent examples include Mahdizadeh Sani et al., 2024; Miletić and Schulte im Walde, 2025; McGillivray et al., 2025; Würschinger and McGillivray, 2025, i.a.).

For German, we train word2vec models from scratch. We use the skip-gram algorithm with negative sampling, setting window size to 5, vector dimensions to 200, minimum frequency to 2, learning rate to 0.03, and we train for 15 epochs. These hyperparameters were selected after a non-exhaustive grid search evaluated on the German semantic relatedness dataset WordSim280 (Köper et al., 2015).

For English, we rely on previously published models for CCOHA and RSC (Zaberer, 2025), which enables us to implement a resource reuse scenario. The models were trained using the continuous bag-of-words (CBOW) algorithm, with minimum frequency set to 1 and default values for the remaining hyperparameters: windows size of 5, vector dimensions of 100, learning rate of 0.025, and training for 5 epochs.

Once they have been trained, individual word2vec models define embedding spaces whose axes have no absolute meaning and as such cannot be directly compared. To address this issue, we apply the Orthogonal Procrustes method, which aligns two embedding spaces by finding a rotation matrix which minimizes the difference between the spaces while preserving distances and angles (Schönemann, 1966; Hamilton et al., 2016). We individually align all German models to the one trained on the last time slice of the Spiegel corpus, and all English models to the one trained on the last time slice of the CCOHA corpus.

3.2. Dimensionality reduction

Core components of our system visualize the words represented by our word embedding models in a two-dimensional space. We therefore need to reduce our aligned word embedding models from 100 or 200 dimensions (for English and German, respectively) into two dimensions. We do so using Pairwise Controlled Manifold Approximation Projection (PaCMAP; Wang et al., 2021), a dimensionality reduction technique which takes into account neighbor, mid-near, and further pairs of points in order to preserve both the global and the local structure of the vector space. This property is highly relevant for our use case: the global structure enables us to identify broad areas of interest in the lexicon, while the local structure is critical when inspecting the semantic neighborhoods of individual words. PaCMAP is also less sensitive to hyperparameter settings than other widely used techniques (e.g., t-SNE; van der Maaten and Hinton, 2008) and as such is less susceptible to spurious differences in visualization outputs.

3.3. Neighborhood density

Our system provides information on the general semantic properties of individual words by computing neighborhood density, which has been previously used to characterize semantic change over time (e.g., Sagi et al., 2009). We implement two variants of neighborhood density suggested by Schulte im Walde and Frassinelli (2022): **TN**, defined as the average cosine score between a target word *t* and its *k* nearest neighbors; and **NN**, defined as the average pairwise cosine score between the *k* nearest neighbors of a target word *t*. When observed over time, an increase in neighborhood density (progressively more related nearest neighbors) reflects semantic specialization; and a decrease in neighborhood density (progressively less related nearest neighbors) reflects semantic generalization. Given a user-selected target word *t*, our system computes the densities on the fly, with the option of interactively adjusting the value of *k*.

4. WoVis Implementation

We now present the core functionalities of WoVis as designed for deployment on a comparison of two contrastive corpora (e.g., corresponding to two language varieties), each of which is further partitioned into temporal subcorpora. As input, WoVis requires the word2vec models trained on individual subcorpora and subsequently aligned to the same vector space; the dimensionality-reduced representations, which can also be obtained by running the corresponding WoVis module; and an optional file containing supplementary lexical features. See Appendix A for further input requirements.

As shown in Figure 1, WoVis produces visualizations at different levels of granularity: a general overview plot, which shows a two-dimensional projection of all words in all subcorpora (panel A); a neighborhood overview plot, which shows a two-dimensional projection of a selected target word and its nearest neighbors (panel B) together with a table listing the neighbors and their linguistic properties (panel C); overviews of neighborhood evolution for a selected target word, shown in terms of their overlap for all pairs of subcorpora (panel D) and neighborhood densities measured within individual subcorpora (panel E); and a word pair plot, showing the change in relatedness of a target word to user-selected comparison words (panel F). The design of the individual visualizations is explained in detail in the remainder of this section, while the descriptive insights they provide are illustrated using the case studies in Section 5.

General overview The general overview (Figure 1, panel A) uses a scatter plot to show all words from each subcorpus, enabling the user to explore the whole vocabulary and detect clusters of interest. The coordinates of individual points correspond to the two-dimensional PaCMAP reduction of the aligned word2vec embeddings. The plot uses marker shapes and colors to distinguish the corpora (e.g., Spiegel vs. Berliner Zeitung) and brightness to indicate time periods. When hovering over a data point, a label shows the corresponding word and subcorpus. The plot supports zooming into specific areas; brushing, i.e., selecting a subset of points on which to focus; and showing only a selection of subcorpora, with the labels in the legend box serving as toggle buttons.

In a typical second step, the user will move onto an individual word, either by clicking on the corresponding point or by looking it up using the provided auto-complete search box. This action will trigger the neighborhood overview.

Neighborhood overview The neighborhood overview (Figure 1, panel B) appears below the general overview and functions in the same way, with the important exception of the displayed data points. It shows the previously selected target word (using a star-shaped marker) and its *k* nearest neighbors in each subcorpus, with the value of *k* set interactively (between 1 and 10,000). This plot enables the user to quickly identify clear differences in meaning across corpora and/or time periods. We also link the two overview plots, showing the cursor location in the neighborhood overview as a crosshair in the general overview; this enables smooth iterations between the two exploration modalities.

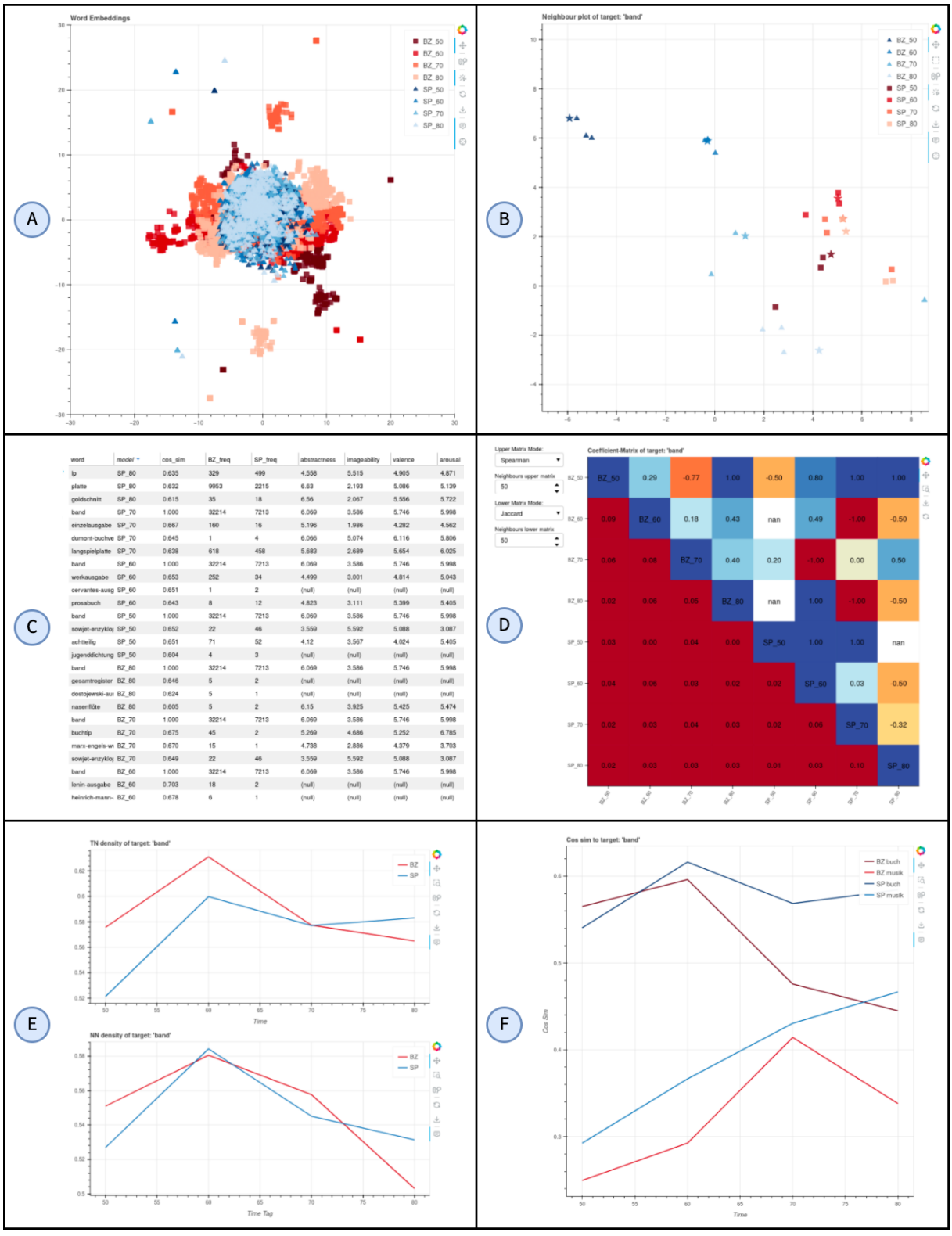


Figure 1: WoVis views for West vs. East German comparison. **A:** general overview; **B:** neighborhood overview; **C:** neighborhood table; **D:** neighborhood overlap over time; **E:** neighborhood density over time; **F:** word pair plot. The shown plots are from our German-language experiments: panels B–E illustrate the neighborhood of the word *Band*; panel F compares it with words *Buch* and *Musik*.

The neighborhood overview also includes a table with the displayed words (Figure 1, panel C), showing the cosine score between the nearest neighbors and the target word as well as any optional model-external lexical features (in our case, frequency and affective norms). A cross-selection between the table and the scatter plot is also supported.

Neighborhood overlap over time As a more structured view of the evolution of a target word's meaning, WoVis provides a heatmap showing the pairwise similarities of the target word's distributional neighborhoods *across* all pairs of subcorpora (Figure 1, panel D). This enables the user to identify pairs of subcorpora with strong shifts in sets of nearest neighbors, which may correspond to points of time in which semantic change occurs.

The neighborhood information is summarized using two measures. The upper triangle of the heatmap shows Spearman's rank-order correlation coefficient, which is calculated based on the cosine scores between the target word and the nearest neighbors; it captures the similarity of nearest neighbors from two corpus sections while accounting for their ordering. The lower triangle of the heatmap shows the Jaccard index, which measures the overlap between two sets of nearest neighbors. This measure is simpler since it ignores the ordering, but potentially more intuitive for the user. The number of nearest neighbors and the position of the two measures in the heatmap can be customized.

Neighborhood density over time To characterize the general semantic properties of a target word *within* a given corpus section, WoVis provides neighborhood density plots (Figure 1, panel E). They show the density value (y-axis) as it evolves over time (x-axis), with line styles used to distinguish the corpora. This view enables the user to identify semantic specialization vs. generalization over time. Plots are provided for both TN and NN density formulations (see Section 3.3), and the number of neighbors can be interactively specified.

The same plot can be toggled to an alternative view showing neighborhood density (y-axis) over the number of neighbors (x-axis), with separate lines representing all subcorpora. This functionality can help determine the useful number of nearest neighbors for a given target word.

Word pair plot As an operational analysis of a target word's dominant senses over time, WoVis provides word pair plots (Figure 1, panel F). These are line plots showing the cosine similarity (y-axis) of the target word to two manually specified reference words, which in a typical scenario reflect two competing senses. The similarity values are plotted over time (x-axis) for each of the two corpora. The user can thus see if the target word has become more or less related to either reference word.

5. Case studies

We now present case studies of semantic change illustrating different insights provided by WoVis.

5.1. Spiegel vs. Berliner Zeitung

We begin with the comparison of German-language corpora: Spiegel, containing texts from West Germany; and Berliner Zeitung, with texts from East Germany. We discuss an example of a lexical process that was highly typical for these two varieties: culture-related English loanwords, which were generally introduced in the West and then spread to East Germany with a delay.

As an example of such a usage, we consider the homonymous word *Band*. In addition to its earliest attested sense 'connection' (neuter; 7th century), it more recently developed the sense 'book volume' (masculine; 18th century) and, borrowing from English, 'musical group' (feminine; 20th century).[4] We mainly focus on the competition between the two more recent senses.

Visualizations for *Band* are shown in Figure 1; we being by observing the neighborhood overview plot (panel B). For Spiegel, the *Band* neighborhood is concentrated in a stable area over the whole time period. For Berliner Zeitung, it is more widespread and moves closer to the Spiegel area over time; this is indicative of progressive convergence. To better understand the senses corresponding to this tendency, we inspect the list of nearest neighbors (Figure 1, panel C; full list in Table 2). In Spiegel, *Band* first appears in music-related contexts in the 1970s, as shown by the nearest neighbor *Langspielplatte* 'long-playing record'. In Berliner Zeitung, music-related nearest neighbors only appear in the 1980s. This indicates a lag in the introduction of the English-borrowed usage in East Germany.

In terms of more general semantic properties, the neighborhood density plot (Figure 1, panel E) shows that density values for both corpora peak in the 1960s and subsequently decrease. The information in this plot subsumes the full range of uses of the target word, so we cannot pinpoint with certainty the most dominant individual sense at a given point in time. However, the overall trend reflects a progressive diversification of contexts in which the target word is attested, which is consistent with the introduction of a new sense after the 1960s.

Bringing together the initial insights, we now set up a word pair plot for the target word *Band* (Figure 1, panel F). We visualize its relatedness

[4] https://www.dwds.de/wb/Band

Spiegel					
1960–1969		1970–1979		1980–1989	
werkausgabe	*collected works*	einzelausgabe	*single issue*	lp	*LP*
cervantes-ausgabe	*Cervantes edition*	dumont-buchverlag	*Dumont Book Publisher*	platte	*record*
prosabuch	*prose book*	langspielplatte	*long-playing record*	goldschnitt	*gilt edge*
silberlöwe	*silver lion*	volkspoesie	*folk poetry*	rockblatt	*rock paper*
subskriptionspreis	*subscription price*	dumont-verlag	*Dumont Publisher*	doppelalbum	*double album*
romanreihe	*novel series*	edition	*edition*	konzert-ausschnitt	*concert excerpt*
einzelausgabe	*single issue*	jubiläumsausgabe	*anniversary edition*	solo-album	*solo album*
werkstattreihe	*workshop series*	originalplatte	*original record*	nachhören	*listen again*
mustervorstellung	*sample presentation*	lp	*LP*	shoes	*shoes*
dtv-band	*dtv volume*	folio	*folio*	gesamtverzeichnis	*full catalog*

Berliner Zeitung					
1960–1969		1970–1979		1980–1989	
lenin-ausgabe	*Lenin edition*	buchtip	*book recommendation*	gesamtregister	*complete index*
heinrich-mann-ausgabe	*Heinrich Mann edition*	marx-engels-werk	*Marx and Engels works*	dostojewski-ausgabe	*Dostoevsky edition*
heine-ausgabe	*Heine edition*	sowjet-enzyklopädie	*Soviet Encyclopedia*	nasenflöte	*nose flute*
werkausgabe	*collected works*	lenin-ausgabe	*Lenin edition*	brennesselsuppe	*nettle soup*
einzelband	*single issue*	urania-universum	*Urania Universum*	beethoven-konzert	*Beethoven concert*
springfrosch	*jumping frog*	wissenschaftsverlag	*scientific publisher*	tanzmusik-sendung	*dance music program*
übersetzung	*translation*	heft	*issue, notebook*	konzertfilm	*concert film*
punktschrift	*Braille script*	essay	*essay*	kunstspiel	*art play*
paperback	*paperback*	goethe-ausgabe	*Goethe edition*	tolstoi-ausgabe	*Tolstoy edition*
edition	*edition*	pierer	*Pierer*	brecht-ausgabe	*Brecht edition*

Table 2: Top 10 nearest neighbors of *Band*, with English translations shown in italics. First decade omitted for space reasons.

(as measured by the cosine score) to two sense-specific reference words: *Buch* 'book', indicative of the 'book volume' sense; and *Musik* 'music', indicative of the 'musical group' sense. In both corpora, the relatedness to *Buch* is highest in the 1960s and subsequently decreases. The relatedness to *Musik* also increases: in the Spiegel corpus, from about 0.30 in the 1950s to 0.45 in the 1980s; and in the Berliner Zeitung corpus, from 0.25 to its highest point at 0.40 in the 1970s, then dropping again to 0.35 in the 1980s. These insights pinpoint the trend of *Band* being increasingly used in music-related contexts in both corpora, with a stronger shift in the West than in the East of Germany.

5.2. RSC vs. CCOHA

We now move on to the comparison of English-language corpora: RSC, which represents scientific UK English; and CCOHA, which represents general-domain US English. Since this is a more complex corpus comparison, we discuss two types of variation: differences in word usage across domains and across country-level language varieties.

5.2.1. Variation across domains

As an example of variation in the scientific vs. general domain, consider the case of *gold*. The WoVis-provided nearest neighbors reflect its dominant sense – referring to the precious metal – in both RSC and CCOHA (Table 3). But there are also differences in the contexts in which that sense is used. In RSC, the top nearest neighbors are limited to a broad set of metals (*zinc*, *aluminum*). This points to the technical nature of the underlying contexts, which is further supported by the hyponym *metal* among the nearest neighbors. In CCOHA, the nearest neighbors include other metals (*silver*, *copper*), but also objects made of gold (*jewels*, *coin*) as well as other precious materials used for decorative purposes (*diamond*, *pearl*). This trend indicates more varied, typically everyday, usage contexts in CCOHA.

The qualitative differences observed thus far are also reflected by the quantitative properties computed by WoVis. Neighborhood densities are consistently higher for the RSC than for CCOHA (Figure 2, top), confirming a more restricted set of usage contexts in the scientific domain. We also examine a word pair plot (Figure 2, bottom), selecting as references *iron* (as a type of metal) and *pearl* (as a decorative material). RSC shows clearly higher relatedness for *iron*, and CCOHA for *pearl*; for both corpora, the cosine values are in the 0.7–0.8 range and stable over time. CCOHA also exhibits a moderately strong and stable relatedness with *iron* ($\approx$0.6), indicating that the more technical usage contexts are also attested throughout the time span of the corpus. By contrast, RSC shows limited relatedness with *pearl*, dropping from $\approx$0.4 to $\approx$0.2. This finding suggests further specialization of the scientific contexts over time.

RSC			
1800–1849	1850–1899	1900–1949	1950–1996
copper	silver	silver	silver
platina	palladium	zinc	copper
silver	copper	tin	aluminium
tin	platinum	aluminium	nickel
platinum	bismuth	copper	platinum
bismuth	antimony	nickel	tantalum
metal	zinc	platinum	beryllium
zinc	metal	cadmium	tungsten
iron	tin	rhodium	tin
steel	sulphur	graphite	iron

CCOHA			
1810–1849	1850–1899	1900–1949	1950–2009
silver	silver	silver	silver
diamonds	copper	copper	diamond
pearls	pearl	bullion	diamonds
pearl	diamond	pearl	copper
jewels	diamonds	diamonds	rubies
rubies	pearls	diamond	pearl
bracelets	coin	rubies	ruby
diamond	rubies	nuggets	jade
copper	nickel	turquoise	lazuli
coin	jewels	pearls	sapphire

Table 3: Top 10 nearest neighbors of *gold*.

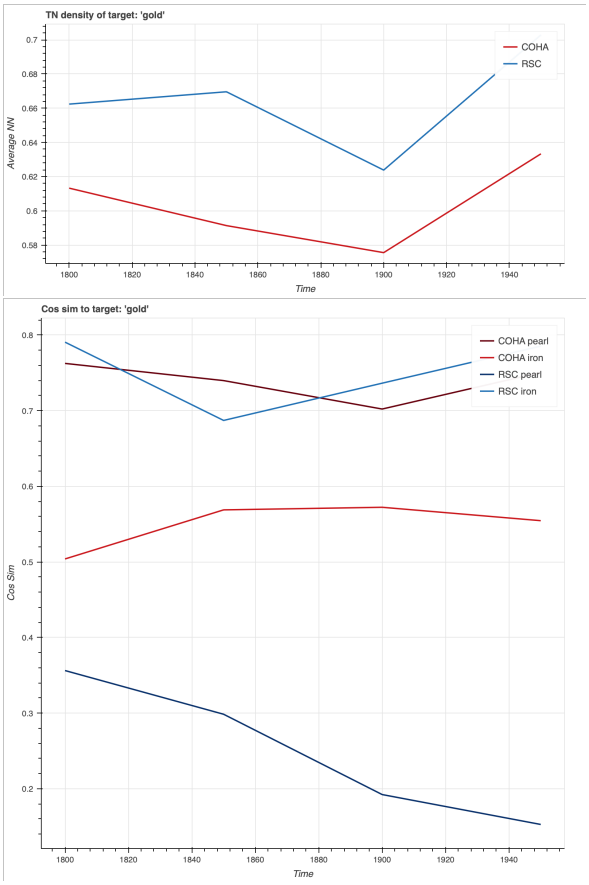


Figure 2: Sample visualizations for *gold* in RSC and CCOHA. Top: TN density (k=100); bottom: word pair plot with references *iron* and *pearl*.

5.2.2. Variation across language varieties

We now look at the ability of our corpora to capture geographic variation in UK vs. US English. We examine the word *gas*, whose dominant sense refers to a state of matter and is attested since the late 18th century.[5] In US (but not UK) English, *gas* is also widely used as an abbreviation of *gasoline*, becoming common by the end of the 19th century.[6]

This characterization finds empirical support in the WoVis lists of nearest neighbors (Table 4). The RSC usage of *gas* corresponds to the earlier sense, as reflected by a diachronically stable set of nearest neighbors: different gases (*hydrogen*, *oxygen*, *chlorine*) as well as other states of matter and related concepts (*liquid*, *fluid*, *air*). CCOHA points to the same nearest neighbors, but importantly also energy sources starting from the 20th century, with prominent nearest neighbors including *gasoline*, *fuel*, and *electricity*.

We once again compare these qualitative differences with the development of neighborhood density (Figure 3, top). CCOHA has a consistently higher neighborhood density, in apparent contrast with the higher degree of polysemy – and hence lower expected density – suggested by individual nearest neighbors. However, it might be the case that the 'gasoline' sense was not only introduced to the US English usage, but also became prevalent to the point of restricting contextual variability. This hypothesis is consistent with the highest density difference between CCOHA and RSC appearing at the beginning of the 20th century, in line with the growing practical significance of gasoline.

We also examine a word pair plot (Figure 3, bottom), selecting as reference words *air*, which references the 'state of matter' sense; and *fuel*, which references the 'gasoline' sense. The plot shows a diachronically stable ordering of dominant senses: the RSC usage is more related to *air* and the CCOHA usage to *fuel*, with the cosine value in both cases ≈0.1–0.3 higher than for the alternative reference word. But beyond this relative difference, both corpora exhibit a diachronic increase in relatedness to *fuel*, consistent with the previously noted differences in neighborhood density.

[5] https://www.etymonline.com/word/gas

[6] https://www.etymonline.com/word/gasoline

RSC			
1800–1849	1850–1899	1900–1949	1950–1996
hydrogen	air	air	liquid
oxygen	gases	gases	gases
chlorine	oxygen	oxygen	air
gases	steam	steam	vapour
vapour	liquid	hydrogen	helium
nitrogen	hydrogen	helium	gaseous
hydrogene	vapour	liquid	reactant
mixture	ozone	nitrogen	solvent
oxygene	argon	vapour	water
phosphorus	chlorine	water	fluid
CCOHA			
1810–1849	1850–1899	1900–1949	1950–2009
acid	carbon	steam	gasoline
fluid	fluid	electricity	fuel
carbonic	boiler	gasoline	electricity
hydrogen	oxygen	helium	steam
oxygen	furnace	oxygen	propane
carbon	vapour	fuel	carbon
liquid	electricity	sulphur	coal
sulphuric	combustion	gases	oxygen
sulphur	wick	alcohol	methane
alcohol	oil	carbon	hydrocarbon

Table 4: Top 10 nearest neighbors of *gas*.

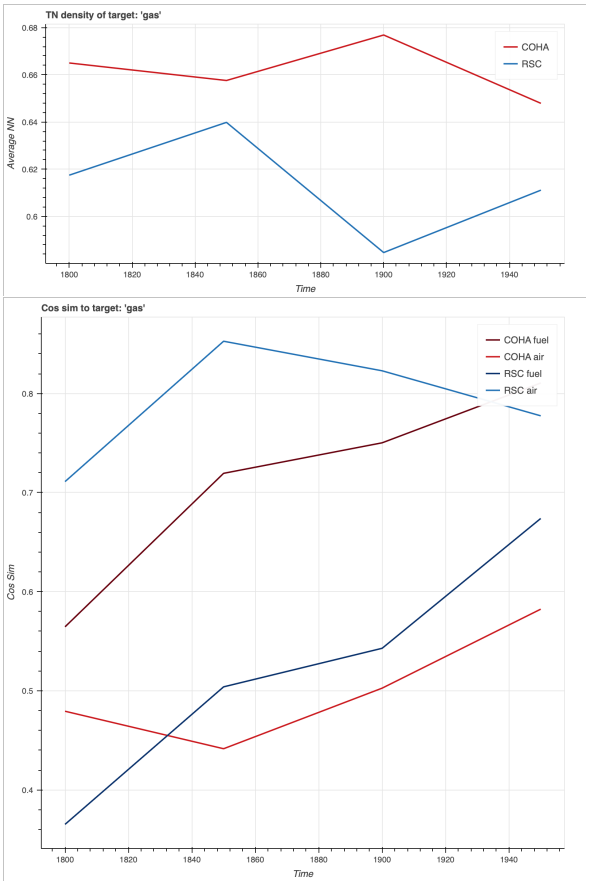


Figure 3: Sample visualizations for *gas* in RSC and CCOHA. Top: TN density (k=100); bottom: word pair plot with references *air* and *fuel*.

6. Conclusion

We have presented WoVis, an interactive visualization tool for comparisons of word embedding models in semantic change analyses. Explicitly designed to foster novel linguistic insights from existing historical and dialectal language resources, our system supports complex model comparisons capturing two dimensions of variation, e.g., in order to simultaneously contrast language use in different language varieties and time periods. We have deployed WoVis in two scenarios corresponding to the dual-comparison configuration: West vs. East Germany, 1950–1989; and general-domain US vs. scientific UK English, ca. 1800–2000. These case studies demonstrate the utility of our system for research into the evolution of non-standard and domain-specific language varieties, and we hope that WoVis will contribute to a more efficient use of the language resources that target them.

7. Limitations

We illustrated the utility of our visualization method on data for German and English. While the datasets we used correspond to the target category of historical and dialectal language resources, both languages are very highly resourced, which also facilitates access to these kinds of datasets. The overall amount of available language data is considerably smaller for lower-resourced languages and language varieties, which may negatively affect the robustness of the word embedding models trained on them and consequently limit the informativeness of the visualizations produced by our system.

Moreover, we principally assessed our system through manual inspection of examples of interest. In this way, we simulated system deployment in linguistic research, observing both practical usability as well as potential descriptive insights provided by the visualizations. This approach could be complemented with additional quantitative evaluations (e.g., relying on clustering quality metrics), as well as a more exhaustive set of possible system components (e.g., different word embedding models, dimensionality reduction techniques, and so forth).

8. Acknowledgments

The work presented here was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Research Grant SCHU 2580/5-2 (*Computational Models of Semantic Variation in Multi-Word Expressions across Speakers and Languages*), under Project-ID 251654672 – TRR 161 (project A08), and under Germany's Excellence Strategy – EXC 2120/1 – 390831618.

9. Bibliographical References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Ulrich Ammon. 1991. The differentiation of the German language into national varieties of the Federal Republic of Germany (F.R.G.), the German Democratic Republic (G.D.R.), Austria and Switzerland. *History of European Ideas*, 13(1-2):75–88.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*, 1 edition. Cambridge University Press.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence*, 3:73.

Andreas Blank. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*, pages 61–90. De Gruyter.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Joan Bybee. 2015. *Language Change*. Cambridge University Press.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 – 12th International Conference on Computational Semantics – Long Papers*.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802, Marseille, France. European Language Resources Association.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Manfred W. Hellmann. 1993. Ostdeutsch - Westdeutsch im kontakt. Brücke oder schranke der verständigung? *Germanistische Mitteilungen*, 38:3–35.

Adam Jatowt, Nina Tahmasebi, and Lars Borin. 2021. Computational approaches to lexical semantic change: Visualization systems and novel applications. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Lexical Semantic Change*. Language Science Press, Berlin.

Christopher Jenkins. 2026. *Compound(ing) Interest in Interesting Compounds or, Semantic Change and Distributional Representation of Noun Compounds*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1928–1931, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Kinne and Birgit Strube-Edelmann. 1981. *Kleines Wörterbuch des DDR-Wortschatzes*, 2. aufl edition. Schwann, Düsseldorf.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and "semantic" structure of continuous word

spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 40–45, London, UK. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož, Slovenia. European Language Resources Association (ELRA).

Samin Mahdizadeh Sani, Malak Rassem, Chris Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2024. What can diachronic contexts and topics tell us about the present-day compositionality of English noun compounds? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17449–17458, Torino, Italia. ELRA and ICCL.

Maximilian Maurer, Chris Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2023. Classifying noun compounds for present-day compositionality: Contributions of diachronic frequency and productivity patterns. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 40–51, Ingolstadt, Germany. Association for Computational Lingustics.

Barbara McGillivray, Nilo Pedrazzini, Arianna Ciula, Jon Lawrence, Tiffany Ong, Mia Ridge, and Jose Miguel Monteiro Vieira. 2025. Analysing the language of mechanisation in nineteenth-century British newspapers. In Ruth Ahnert, Emma Griffin, and Jon Lawrence, editors, *Living with Machines*. University of London Press.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Filip Miletić, Anne Przewozny-Desriaux, and Ludovic Tanguy. 2021. Detecting contact-induced semantic shifts: What can embedding-based methods do in practice? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10852–10865, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Filip Miletić and Sabine Schulte im Walde. 2025. Modeling the evolution of English noun compounds with feature-rich diachronic compositionality prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20071–20092, Vienna, Austria. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Malak Rassem, Myrto Tsigkouli, Chris W. Jenkins, Filip Miletić, and Sabine Schulte im Walde. 2024. Visualising changes in semantic neighbourhoods of English noun compounds over time. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 240–246, Miami, USA. Association for Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23. International Committee for Computational Linguistics.

Peter H. Schönemann. 1966. A generalized solution of the Orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

Sabine Schulte im Walde and Diego Frassinelli. 2022. Distributional measures of semantic abstraction. *Frontiers in Artificial Intelligence*, 4:796756.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2018. Inducing a lexicon of sociolinguistic variables from code-mixed text. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages

1–6, Brussels, Belgium. Association for Computational Linguistics.

Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2016. Embedding Projector: Interactive visualization and interpretation of embeddings. In *Proceedings of NIPS 2016 Workshop on Interpretable Machine Learning for Complex Systems*. arXiv.

Claudia Soria, Joseph J Mariani, and Carlo Zoli. 2013. Dwarfs sitting on the giants' shoulders - how LTs for regional and minority languages can benefit from piggybacking major languages. In *Conference of the Foundation for Endangered Languages*, Ottawa, Canada.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data-mining approach using register features. *Journal of the Association for Information Science and Technology*, 67(7):1668–1678.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

Quirin Würschinger and Barbara McGillivray. 2025. Semantic change and socio-semantic variation: The case of COVID-related neologisms on Reddit. *Linguistics Vanguard*, 11(s3):339–352.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 673–681, New York, NY, USA. Association for Computing Machinery.

Urs Zaberer. 2025. Noun compound compositionality in scientific English. Master's thesis, University of Stuttgart, Stuttgart, Germany.

A. Input Format

For each subcorpus being visualized (e.g., Berliner Zeitung 50s, Berliner Zeitung 60s, etc.), WoVis requires the following input files:

- A trained word2vec model in the Gensim KeyedVectors format together with the corresponding NumPy file;
- The corresponding two-dimensional word and coordinate NumPy file, which can also be produced by running the requisite WoVis module;
- An optional CSV file containing model-external lexical features (e.g., word frequency and affective norms), with the first column containing the target words, and the header containing the labels for the included information.

All file names consist of two components: a corpus name tag and a time tag, separated by an underscore (e.g., BZ_50). This information is used when rendering the plots. The name tag determines the general base color in all plots and the marker shape in scatter plots, while the time tag controls the brightness, with more recent data represented using lighter shades. Further implementation details are provided in WoVis documentation.

Speaker Normalization via Voice Conversion Reveals a Human–Machine Dissociation in Dialect Classification

Caroline Kleen[1,*], **Lea Fischbach**[1,*], **Akbar Karimi**[2,3], **Lucie Flek**[2,3], **Alfred Lameli**[1]
[1]Research Center Deutscher Sprachatlas, Marburg University, Germany
[2]Lamarr Institute for ML and AI, Germany
[3]b-it Center, University of Bonn, Germany
{caroline.kleen, lea.fischbach}@uni-marburg.de, {akbar.karimi, lucie.flek}@uni-bonn.de, lameli@uni-marburg.de
*Equal contribution

Abstract

This study evaluates whether Retrieval-based Voice Conversion (RVC) can be used to normalize speaker-specific variability while preserving dialect-relevant acoustic cues, and what the response of human and machine systems to this manipulation reveals about the architecture of dialect recognition. In two perception experiments, speech samples from nine German dialect regions were presented either in their original form or after conversion to a single target speaker. We compared overall accuracy, confusion structures, item-level response distributions, and the interaction between listener origin and target dialect across conditions. Human classification remained stable under voice conversion. Accuracy did not differ between conditions, confusion matrices were highly correlated, and item-level divergences were minimal. The interaction between listener origin and target dialect—reflecting systematic regional bias—remained invariant. These findings indicate that RVC does not distort perceptually relevant dialectal cues and that human dialect recognition is robust to speaker normalization. In contrast, we evaluated a deep learning model under matched conditions: model accuracy improved significantly under RVC, while human performance remained unchanged. This dissociation reframes RVC as an experimental probe for investigating the divergence between human and machine speech processing, suggesting that this divergence is rooted in fundamentally different representational architectures.

Keywords: Retrieval-based voice conversion, speaker normalization, dialect classification, human perception, deep learning

1. Introduction

A defining feature of human speech perception is its robustness to inter-speaker variability. Listeners effortlessly recognize the same linguistic categories across speakers differing in age, sex, regional background, and vocal tract morphology. This is a capacity referred to as *speaker normalization* (Johnson and Sjerps, 2021). Crucially, this normalization does not operate on absolute acoustic values but encodes linguistic categories relative to speaker-specific context (Sjerps et al., 2019). Dialect recognition represents a particularly demanding instance of this capacity. Listeners must simultaneously normalize for speaker-level variability and attend to dialect-level variation, which is itself encoded in the same acoustic dimensions that differ across speakers.

Deep learning models for dialect classification face an analogous challenge, but address it through an entirely different mechanism. Rather than deploying pre-structured normalization routines shaped by years of linguistic exposure, neural models optimize directly on the statistical structure of the training data. As a consequence, their internal representations may conflate speaker-specific and dialect-specific variance whenever these co-occur in the training signal. This problem is well-known in low-resource settings where limited speaker diversity can lead models to overfit to speaker characteristics rather than linguistic variation (Fischbach et al., 2025a).

Retrieval-based Voice Conversion (RVC) offers a way to disentangle these sources of variance computationally. By mapping utterances from multiple speakers onto a single target voice, RVC attenuates speaker-specific acoustic properties while preserving the linguistic signal (Sisman et al., 2020). Previous work has shown that this procedure improves dialect classification performance in low-resource settings (Fischbach et al., 2025a) and geographic regression tasks (Gutscher and Pucher, 2025). However, these findings address the technical validity of RVC as an augmentation method. It remains unclear whether RVC constitutes a valid normalization technique for dialectal speech resources beyond performance gains. A different question is what RVC reveals about the architecture of dialect recognition itself. These two questions motivate the present study.

We address this question by adopting a parallel evaluation design in which the same stimulus material—original and RVC-converted dialectal speech from nine German dialect regions—is presented to both human listeners and a neural classifier under matched task conditions. We hy-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 177–187
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

pothesize that if human and machine dialect recognition rely on the same acoustic representations, both should respond similarly to RVC. If they rely on different representations, their responses should diverge. We exploit this logic to investigate whether speaker normalization by RVC is compatible with the cognitive normalization mechanisms that human listeners deploy. The results reveal a sharp dissociation. RVC systematically improves machine classification while leaving human performance entirely unaffected. We argue that this dissociation is not incidental but reflects a fundamental difference in how biological and artificial systems represent dialectal speech.

The remainder of the paper is organized as follows. Section 2 presents the material and methods: corpus and stimuli, the acoustic characterization of RVC, and the evaluation design including both the computational and perceptual experiments. Section 3 presents the results for human and model classification and quantifies the dissociation directly. Section 4 discusses theoretical implications, and Section 5 concludes.

2. Material and Methods

2.1. Data

The stimuli presented to human listeners are randomly sampled from the same pool of recordings used for model training and evaluation (the REDE corpus (Schmidt et al., 2020–)), supporting comparability across conditions, although the exact stimulus sets differ between the two evaluation setups. The REDE corpus comprises recordings from male speakers[1], from three age groups throughout Germany: young (17–26 years), middle-aged (42–59 years) and older speakers (60+ years), captured in five to six recording situations[2]. For the purpose of this study, only data from the older age group is used, as this group has been shown to exhibit particularly high dialect competence (Lameli, 2025), providing stimuli that maximize the signal-to-noise ratio for dialect-relevant acoustic features, which is a prerequisite for a clean experimental test of normalization robustness. Furthermore, only the recording situation of the so-called dialectal "Wenker-Sentences" (Wenker, 2013) are analyzed. In this task, the participants translated 40 standard German sentences, spoken by the interviewer, into their local dialect.

The recordings are first automatically segmented using a speaker diarization pipeline to identify stretches of speech belonging to the target speaker; details of this procedure are reported in a companion paper (Fischbach, 2024). These excerpts are subsequently concatenated per speaker and divided into consecutive 10-second segments to obtain multiple samples per speaker of equal duration. The choice of 10 seconds is motivated by preliminary experiments indicating that shorter durations lead to noticeably lower classification performance by a Deep Learning (DL)-model, whereas longer segments did not result in significant performance gains. Ten-second segments are also a practical duration for perceptual evaluation: they provide sufficient phonological material for dialect recognition while keeping listening fatigue within acceptable bounds for an online perception study. All audio files are pre-processed by converting them to mono, setting the bit depth to 16 bits, and the sampling rate to 16 kHz, which matches with the specifications of the DL-model used for this experiment.

The classification tasks consider a total of ten distinct German dialect areas, categorized according to Lameli (2013). Due to insufficient data, the Frisian language area was excluded from the experiments, resulting in nine dialect classes. Figure 1 illustrates the spatial distribution of these nine dialect areas across Germany; the excluded Frisian area is located in the northwestern corner of the map.



Figure 1: Dialect classification of Germany according to Lameli (2013), serving as basis for the classification tasks.

Table 1 provides an overview of these dialect areas, including the number of 10-second samples and speakers used for both the DL-experiment and the human perception study. The first two columns report the total number of generated segments and the corresponding number of speakers in the DL setup, where all available excerpts are included, resulting in a naturally imbalanced

[1]Only male speakers are included, as the REDE project focused on this group for methodological consistency.

[2]Further specifications can be found here: https://rede-infothek.dsa.info/

class distribution across dialect areas. The third column shows the subset selected for the human study. As presenting the full set of segments would have been impractical for perceptual evaluation, the study was restricted to 100 samples in total. The relative class proportions of the DL dataset were retained to ensure comparability between model-based and human evaluation. Within each dialect area, only one sample per speaker was selected in order to maximize speaker diversity. Consequently, in the human study, the number of samples equals the number of speakers per dialect area. In total, the dataset comprises recordings from 166 speakers, amounting to 14.36h of audio data. The recordings originate from 91 different locations in Germany, corresponding to an average of approximately 18.44 speakers per dialect class.

Dialect area	#Samples (DL)	#Speaker (DL)	#Samples (Study)
Bavarian	387	15	8
Brandenburgish	167	4	3
East Central German	462	14	9
East Franconian	512	12	9
Northern Low German	916	34	19
Southern Low German	733	22	14
West Central German	847	30	17
West German	348	10	6
West Upper German	798	25	15
Total	5170	166	100

Table 1: Distribution of 10-second speech segments across dialect areas for the deep learning (DL) experiment and the human perception study (Study). The DL setup includes all available segments. For the perceptual study, a proportional subset of 100 segments was selected. In the human study, the number of samples equals the number of speakers per dialect area.

2.2. Retrieval-based Voice Conversion

Voice conversion is performed using the widely adopted RVCv2 model[3], which standardizes speaker-related acoustic cues by converting all samples to the voice of a single target speaker. As target voice, we employ the publicly available pre-trained weights corresponding to the Jonathan Frakes (German "Beyond Belief: Fact or Fiction" Mix) RVC voice model[4], where *Mix* refers to the use of recordings from multiple German dubbing actors during training. This target voice acoustically corresponds to a male adult speaker of middle age. Previous work has shown that the choice of target speaker, including differences in age groups, does not significantly affect performance when using RVC-based audios for dialect classification, even when multiple or age-matched target voices are employed (Fischbach et al., 2025a).

To evaluate the acoustic impact of the voice conversion process, we analyzed three key acoustic features before and after applying Retrieval-based Voice Conversion (RVC): mean pitch, articulation rate (AR) and the first three formant frequencies (F1, F2, F3). All acoustic features were extracted using Praat (Boersma and Weenink, 2022), executed via Parselmouth (Jadoul et al., 2018) in Python, including the use of the VocalToolkit plugin (Corretge, 2012–2024) for the AR. Since all acoustic features are time-varying, mean values were computed per 10-second segment. The final reported values reflect the mean and standard deviation of these segment-level means. The mean pitch shows only a minor change between the original (118.55 ± 10.71 Hz) and RVC version (118.10 ± 10.68 Hz), with a nearly identical pitch range (59.72 vs. 59.71 Hz). The AR remains similarly stable (original: 4.39 ± 0.49 syl/sec; RVC: 4.42 ± 0.46 syl/sec), indicating preserved intonational and temporal structure. In contrast, the long-term average formant (LTAF) frequencies are more affected, although both the difference between the original and RVC and the variation range (see error bars in Figure 2) are negligible in phonetical and phonological terms. Also, the variation is within the allophonic range, as evidenced by Sendlmeier and Seebode (2006) for the vowels of German: F1 increases slightly (from 536.20 ± 53.38 Hz to 625.22 ± 48.06 Hz), F2 decreases slightly (from 1625.51 ± 89.27 Hz to 1597.58 ± 72.89 Hz), and F3 remains relatively stable (from 2616.68 ± 92.17 Hz to 2667.59 ± 41.75 Hz).

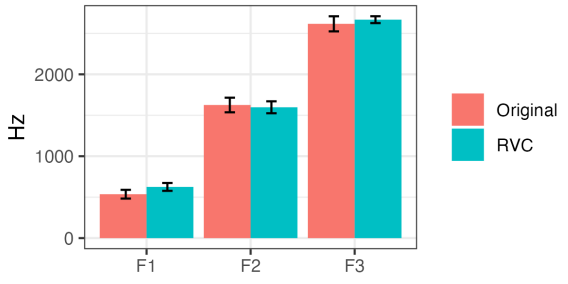


Figure 2: Mean long-term average formant frequencies (F1–F3) before and after voice conversion. Bars represent the mean of segment-level averages (10 s), and error bars indicate the corresponding standard deviation across segments.

These acoustic findings carry different implications for human and machine processing. The sta-

[3] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

[4] https://huggingface.co/Connum/RVC-models/resolve/main/frakes_xfactor_ger.zip

bility of F0 and AR suggests that suprasegmental features relevant to both dialect perception and model classification are preserved. The shifts in F1–F3, by contrast, indicate that RVC primarily targets the spectral envelope, which is the primary acoustic correlate of perceived vocal identity (Fant, 1971). For human listeners, this transformation may be largely inconsequential: research on talker normalization suggests that listeners encode linguistic categories relative to speaker-specific context rather than in absolute acoustic terms (Sjerps et al., 2019), implying that a shift in the overall formant space should not distort within-speaker contrasts that carry dialectal information. For a neural model, however, formant-level shifts may have more complex consequences: they reduce between-speaker variance in the training signal, potentially making dialect-specific patterns more salient. This asymmetry generates a testable prediction that RVC should benefit model performance while leaving human performance unaffected. The following evaluation—which presents comparable stimuli to both human listeners and a neural classifier—directly tests this prediction by comparing the response of each system to the same acoustic manipulation.

2.3. Experiments

To test the predicted dissociation between human and machine dialect recognition under RVC, we adopt a parallel evaluation design in which both systems are exposed to stimuli drawn from the same underlying corpus and perform the same dialect classification task. Rather than a strictly item-matched setup, the two evaluations are designed to be comparable at the level of task, label space, and data distribution, while respecting the methodological constraints of each system.

Specifically, the human study is based on a fixed subset of 100 stimuli (one per speaker) to avoid speaker repetition and perceptual learning effects, whereas the model is evaluated on the full dataset using repeated speaker-disjunct train–validation–test splits. This ensures that the model does not memorize speaker-specific characteristics and provides a robust estimate of generalization performance across the data distribution.

As a result, the comparison should be interpreted as a parallel evaluation of two systems under comparable task conditions, rather than a strictly item-by-item comparison. This design allows us to attribute differences in performance patterns to differences in processing strategies, while accounting for the distinct constraints of human perception and machine learning.

2.3.1. RVC-Augmentation for Deep Learning

The audio material described above serves as input for Google's TRILLsson model (Shor and Venugopalan, 2022)[5], which extracts a vector representation (embedding) for each audio segment. These embeddings are then used to train and evaluate a lightweight Multilayer Perceptron (MLP) consisting of two hidden layers and an output layer with 297,876 trainable parameters. To prevent the model from memorizing speaker-specific characteristics, we apply a strict speaker-disjunct train–validation–test split, ensuring that each speaker appears in only one partition. For each dialect D with $|S_D|$ speakers, we randomly select $\lceil |S_D|/10 \rceil$ speakers for the validation set and $\lceil |S_D|/10 \rceil$ speakers for the test set (the remaining speakers form the training set). Because speaker selection directly affects performance and to get a more reliable estimate of the model's real-world performance, we repeat this procedure across 250 independent runs with different randomly sampled subsets for each run[6].

In the baseline condition, the model is trained exclusively on the original recordings. In the augmented condition, RVC-converted versions of the training samples are added to the training data, while the speaker-disjunct partitioning is maintained. Without RVC-augmentation, the MLP achieves a classification accuracy—defined as the proportion of correctly predicted samples—of 52.7% ($SD = 7.4\%$), whereas with RVC-augmentation the accuracy is increased to 57.2% ($SD = 6.7\%$). This improvement is statistically significant, $t(498) = 7.26$, $p < .001$, $d = 0.65$, based on comparing the 250 baseline runs to the 250 RVC-augmented runs as independent samples. This condition serves as the machine-side component of the parallel evaluation, providing a performance baseline against which the predicted benefit of RVC can be assessed.

2.3.2. RVC in Human Perception

The human perception studies constitute the second component of the parallel evaluation. Critically, the stimuli presented to listeners are drawn from the same pool as those used for model training and testing, ensuring direct comparability of conditions. To contextualize the classification performance of the model presented above, we conducted two perception studies with linguistic laypeople (Anders, 2010),

[5] TRILLsson was selected because it was developed for paralinguistic classification tasks and provides a general-purpose acoustic embedding space without dialect-specific supervision.

[6] The complete pipeline is available at: https://github.com/WoLFi22/DialectClassificationPipeline

which are defined as people without knowledge of linguistics (Klein, 2021). A between-subjects design was adopted to prevent carry-over effects: listeners who first classify original recordings might develop perceptual strategies that transfer to the converted condition, confounding the comparison.

Related work has shown that the design of such tasks strongly affects participants' performance: Hundt et al. (2015) reported accuracies of about 64% when lay participants selected from predefined location options, whereas Kleen (2022) found considerably lower classification performance of 38% in a map-based task without predefined choices. Building on these findings, our perception studies employ a format that minimizes external guidance while still allowing for meaningful geographical categorization, thereby ensuring conceptual comparability with the model, which is likewise evaluated on dialect area categories rather than fixed locations. The map shown to participants, depicting these dialect areas, is presented in Figure 1.

The first study (**Study-O**) investigates how participants classify the original recordings from the corpus, while the second study (**Study-VC**) examines the participants' ratings of the RVC-modified samples. This allows for a direct comparison between human listeners and the model. The surveys were implemented using Gorilla Experiment Builder (`www.gorilla.sc`) to create and host the experiment (Anwyl-Irvine et al., 2020) and distributed using Prolific (`www.prolific.com`) (Prolific, 2024). A total of 100 native German (L1) participants were each presented with all 100 stimuli (10 seconds each) in individually randomized order. The study thus comprised 100 trials per participant, with an estimated completion time of approximately 20–30 minutes including response time and attention checks. Using a geographical map, participants were asked to assign each stimulus to one of ten dialect classes (e.g., Bavarian), with an additional "don't know" option to avoid forced-choice effects in cases of high uncertainty. Although Frisian was included as a response category, no Frisian stimuli were presented. To ensure a direct comparison with the model's forced-choice classification, responses selecting either Frisian or "don't know" were excluded from the subsequent accuracy analyses. As shown in Table 1, the distribution of the 100 stimuli across dialects follows the distribution of the original corpus, and consequently also reflects the number of available speakers per dialect. Importantly, each stimulus in the study originates from a different speaker, ensuring that listeners never evaluated two samples from the same speaker. Participants' engagement was monitored using periodic attention checks (one after every 20 trials). In these trials, no audio stimulus was presented; instead, participants were instructed to select a specific category (e.g., "Please select Bavarian for this trial") to verify they were still attending to the task. Responses to these attention checks were used solely for participant screening and were excluded from the dialect classification analysis.

In Study-O, 50 participants (31 male, 18 female, 1 non-binary; $M_{\text{age}} = 30.06$, $SD = 5.20$ years) achieved an overall classification accuracy of 31.66% ($SD = 13.51\%$). In Study-VC, 50 participants (32 male, 17 female, 1 non-binary; $M_{\text{age}} = 30.06$, $SD = 5.02$ years) achieved an accuracy of 31.38% ($SD = 13.87\%$), statistically indistinguishable from the original condition. A two-sample t-test, based on each participant's total score, did not reveal significant differences in classification performance between the two groups, $t(98) = 0.10$, $p = .919$. A formal equivalence test is reported in Section 3.1.

Participants in both studies were geographically distributed across Germany, ensuring broad regional representation. The regional distributions were highly similar across the two participant groups.

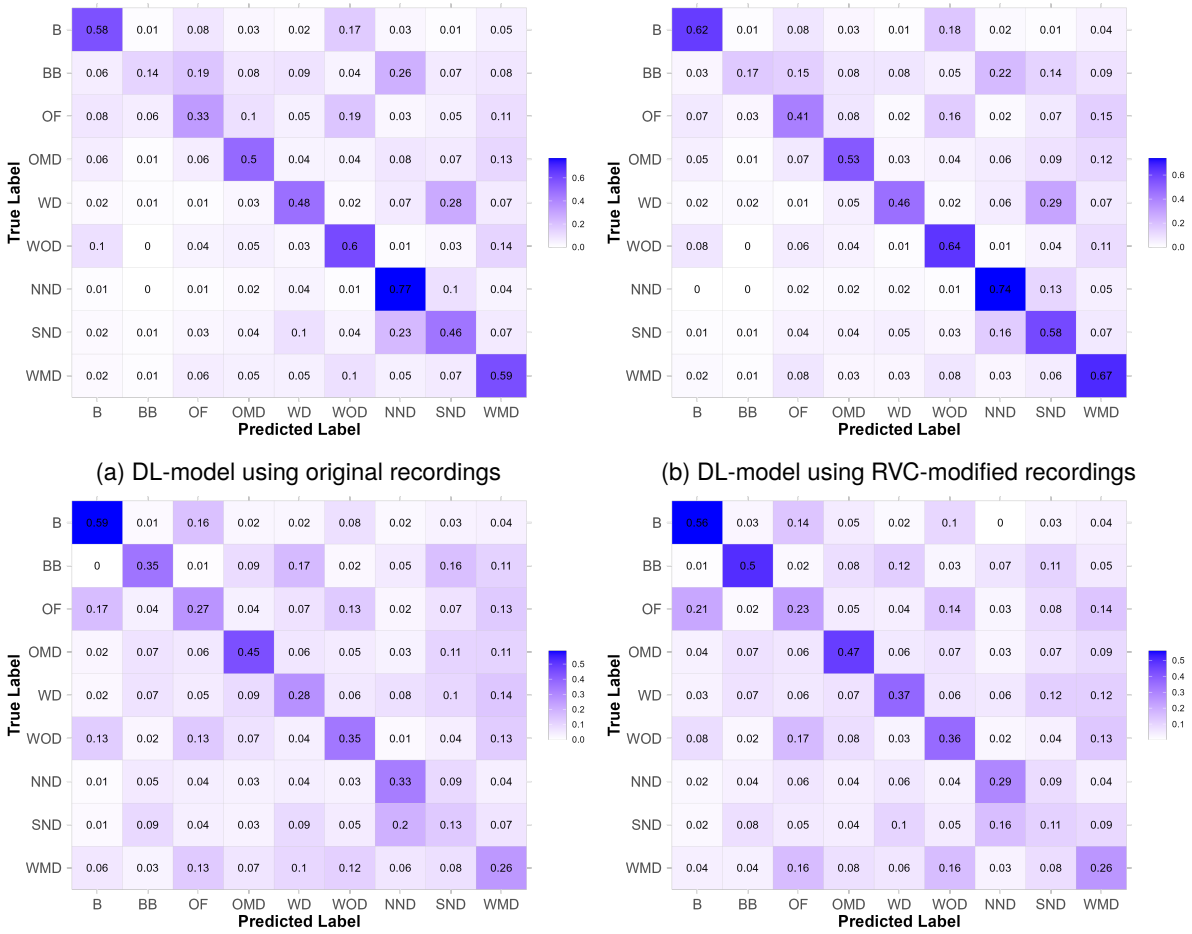
3. Results

We report results in three steps: first, the overall effect of voice conversion on human classification and the modulating role of listeners' geographical origin; second, the model results and third, the direct quantification of the human–machine dissociation.

3.1. Human Classification

3.1.1. Effect of Voice Conversion on human perception

Consistent with the prediction derived from talker normalization theory, human dialect classification showed no response to voice conversion across any of the metrics examined (Study-O and Study-VC). Listener accuracy remained stable (31.66% vs. 31.38%). A Pearson correlation of $r = 0.96$ between the confusion matrices (Figure 3) confirms that error patterns were highly consistent across conditions. We conducted a per-dialect F1-score analysis to ensure that the voice conversion preserves the phonetic dialectal markers. The human performance remains remarkably stable across conditions: while the dialect classes Brandenburgish (+0.102), West German (+0.083) and Northern Low German (+0.052) showed a slight increase in F1-scores, results for Bavarian (–0.015), East Central German (±0.00), West Central German (–0.003), Southern Low German (–0.011) and West Upper German (–0.007) were nearly identical. Only East Franconian showed a marginal decrease (–0.049). These findings confirm that



(a) DL-model using original recordings

(b) DL-model using RVC-modified recordings

(c) Humans using original recordings

(d) Humans using RVC-modified recordings

Figure 3: Confusion matrices showing classification frequencies for the DL-model (a, b) and human listeners (c, d), before and after RVC augmentation. Rows indicate true labels, columns predicted ones. Abbreviations: B = Bavarian; BB = Brandenburgish; OF = East Franconian; OMD = East Central German; WD = West German; WOD = West Upper German; NND = Northern Low German; SND = Southern Low German; WMD = West Central German.

RVC-generated stimuli remain perceptually valid and representative of the target dialects.

To assess whether this stability also holds at the level of individual items, we computed the Jensen-Shannon divergence between the response distributions for each stimulus across conditions. The divergence was consistently low ($M = .091$, $SD = .037$), indicating that the same items were perceived similarly, regardless of whether the original or converted version was presented. Together, these results demonstrate invariance across global performance, confusion structure, and item-level perception across conditions.

To move beyond the absence of a significant difference and provide positive evidence for equivalence, we conducted a two one-sided tests (TOST) equivalence analysis (Lakens, 2017), with equivalence bounds set at ± 5 percentage points, which is a margin we consider the smallest difference of practical significance given the overall accuracy range observed. The analysis confirmed statistical equivalence between the two conditions, $t(98) = -1.72, p = .044$, establishing that human performance under RVC falls within the range of the original condition not merely by failure to reject the null, but by positive equivalence criterion.

3.1.2. Effect of listeners' geographical origin

Beyond general confusion patterns, we investigated whether the normalization process affects regional perception and how classification performance is moderated by the listeners' background. As illustrated in the heatmaps (Figure 4), accuracy is systematically influenced by the geographical origin of the participants. To test the stability of this effect across conditions, we fitted a linear mixed-effects model including a triple interaction between listener origin, target dialect, and study condition (original vs. VC). The analysis revealed a significant main

effect of listener origin, $F(8, 84.3) = 2.37, p = .024$, suggesting that participants' general classification ability differs slightly depending on their own regional origin. In contrast, the main effect of study condition was not significant, $F(1, 84.3) = 0.01, p = .916$, confirming that overall accuracy remained identical between original and voice-converted stimuli. Most importantly, we found a highly significant interaction between listener origin and target dialect, $F(64, 8603.4) = 4.81, p < .001$. Crucially, the triple interaction with study condition was not significant, $F(64, 8603.4) = 1.19, p = .143$. This statistically confirms that the "regional bias" remains invariant across conditions: RVC-based speaker normalization does not alter the cognitive strategies or regional cues listeners rely on.

These results provide a robust explanation for the observed human performance: while accuracy is influenced by both the listener's origin and its interaction with the target dialect, it remains unaffected by the voice conversion process.

3.2. Model Classification

In contrast to the invariant human performance, the deep learning classifier responded systematically to voice conversion. Accuracy increased from 52.7% ($SD = 7.4\%$) to 57.2% ($SD = 6.7\%$), a statistically significant improvement, $t(498) = 7.26, p < .001, d = 0.65$. While the confusion matrices remained highly correlated across conditions ($r = 0.99$), indicating that the overall classification structure was preserved, the consistent accuracy gain across 250 independent runs rules out sampling artifacts and confirms a genuine benefit of speaker normalization for the model.

To control for the potential confound between RVC-specific effects and increased training data quantity, we conducted an additional size-matched augmentation experiment. Specifically, we applied frequency masking to each training sample, following the augmentation strategy proposed by Fischbach et al. (2025b), which was shown to be the most effective augmentation method for dialect classification in prior work. This ensures that the total number of training instances is identical to the RVC-augmented condition. This non-RVC augmentation yields a mean classification accuracy of 55.6% ($SD = 6.3\%$). A direct comparison between the frequency-masked and RVC-augmented conditions shows a statistically significant difference, $t(498) = 2.28$, $p = .023, d = 0.20$, indicating a small but reliable advantage of RVC over generic augmentation. These results suggest that while part of the performance gain can be attributed to increased training data, RVC provides an additional benefit beyond standard augmentation.

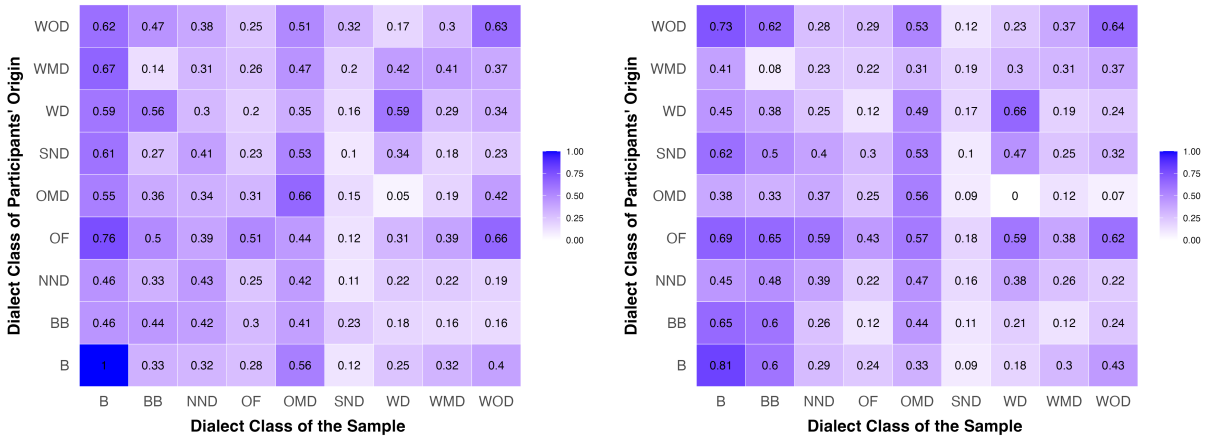
To assess whether the observed improvements are consistent across dialect classes, we additionally conducted a per-dialect analysis using one-vs-rest F1-scores, which provide a more class-sensitive evaluation under dataset imbalance than overall accuracy. The results show consistent improvements for all dialects, with gains ranging from +0.019 to +0.073. Importantly, these improvements are observed both for low-resource dialects (e.g., Brandenburgish, 4 speakers, +0.053) and high-resource dialects (e.g., West Central German, 30 speakers, +0.044), suggesting that the effect is not driven by a subset of dominant classes. Overall, this analysis indicates that the benefit of RVC generalizes across dialect classes and is not an artifact of dataset imbalance.

Notably, the confusion structure of the model differs substantially from that of human listeners even in the baseline condition (see Figures 3a and 3c), suggesting that the two systems do not rely on the same discriminative features. Section 3.3 quantifies this cross-system divergence directly.

3.3. Dissociation

Taken together, the results reveal a double dissociation between human and model behavior under voice conversion. Human accuracy remained stable (31.66% vs. 31.38%), while model accuracy improved (52.7% vs. 57.2%). To quantify this dissociation, we compare condition effect sizes across systems rather than aggregating incommensurable accuracy scales. For the model, RVC produced a gain of 4.5 percentage points, confirmed as statistically significant with a large effect ($d = 0.65, p < .001$, based on 250 independent runs). For human listeners, the corresponding difference was -0.28 percentage points, statistically indistinguishable from zero ($t(98) = 0.10, p = .919$) and confirmed as equivalent to zero by formal TOST criterion (Section 3.1.2). The contrast between the model's effect size ($d = 0.65$) and the near-zero effect for human listeners ($d = 0.02$) directly indexes the magnitude of the dissociation. That the model benefits substantially from a manipulation to which human listeners are entirely insensitive cannot be attributed to differences in task difficulty, stimulus quality, or sampling variability.

A second dimension of the dissociation concerns the confusion structure. While human confusion matrices were highly correlated across conditions ($r = 0.96$), and model confusion matrices similarly stable ($r = 0.99$), the cross-system correlation between human and model confusions was substantially lower ($r = 0.80$, computed as the Pearson correlation between corresponding cells of the two normalized confusion matrices). This indicates that humans and the model not only respond differently to RVC, but classify dialects differently in the first place. This is a finding that bears directly on the interpretation of what each system has learned.



(a) Human classification accuracy by participant origin (Study-O)

(b) Human classification accuracy by participant origin (Study-VC)

Figure 4: Heatmaps showing classification accuracy for the human listeners in Study-O (left) and Study-VC (right). Abbreviations: B = Bavarian; BB = Brandenburgish; OF = East Franconian; OMD = East Central German; WD = West German; WOD = West Upper German; NND = Northern Low German; SND = Southern Low German; WMD = West Central German.

4. Discussion

The central finding of this study is not that RVC improves model performance, nor that human dialect recognition is robust; both results are consistent with prior expectations. The theoretically consequential finding is the combination, namely that the same acoustic manipulation produces opposite responses in the two systems. This double dissociation cannot be explained by differences in task difficulty, stimulus quality, or chance. It reflects a structural difference in how dialectal speech is represented.

The performance gap between human lay listeners (31.66%) and the neural model (52.7%) reflects a fundamental difference in classification strategies. While the model leverages high-dimensional statistical regularities across the entire frequency spectrum, humans rely on a limited set of "shibboleths" or stereotypical markers (e.g., Bavarian features) shaped by media exposure and personal mobility (Köster et al., 2012; Sauer and Hoffmeister, 2022; Hundt et al., 2015). In brief 10-second segments, these prototypical markers may be difficult to detect, forcing lay listeners to rely on subjective mental maps. For human listeners, the results are consistent with exemplar-based models of speech perception (Goldinger, 1998; Johnson, 1997), in which stored representations include speaker-specific information that listeners use as a normalization baseline. Under this account, converting all speakers to the same voice does not help listeners—they already normalize implicitly—but it also does not hurt them, because the normalization baseline adjusts to the new voice. As detailed below, this account receives particularly strong support from the invariance of the regional bias across conditions. This finding goes beyond overall accuracy to the fine-grained structure of dialectal competence.

One alternative interpretation is that human accuracy (31.66%) was too variable or too far from ceiling to detect a subtle RVC effect. However, the TOST analysis rules this out by establishing positive equivalence rather than mere non-significance, and the item-level stability (JSD = 0.091) confirms that individual stimuli were not systematically revalued.

The invariance of the regional bias across conditions provides a particularly stringent test of this account. The interaction between listener origin and target dialect—reflecting the finding that listeners classify nearby dialects more accurately than distant ones (Hundt et al., 2015)—was fully preserved under RVC. This result goes beyond showing that overall accuracy is unaffected: it demonstrates that the fine-grained structure of dialectal competence, shaped by each listener's individual exposure history, is equally unaffected. Under a speaker normalization account, this is precisely what one would expect: if listeners encode dialectal categories relative to speaker context, then shifting the speaker context uniformly—as RVC does—should leave the relative structure of dialectal knowledge intact. The data confirm this prediction at the level of individual listener-dialect interactions.

For the neural classifier, the improvement under RVC is interpretable within a straightforward signal-processing framework: by reducing between-speaker variance, RVC makes dialect-specific patterns more consistently accessible to the model across training instances. This is the mechanism proposed by Fischbach et al. (2025a). What the

present study adds is the contrastive context: the model's sensitivity to speaker-level variance is not a feature of dialect classification per se, but a consequence of the model's architecture. A system with genuine normalization capacity—such as a human listener—should be indifferent to this manipulation. The model is not.

This has practical implications for the use of RVC as an augmentation technique. The human data establish that RVC does not distort dialectally relevant information, providing the perceptual validity that model improvements alone cannot guarantee. RVC-augmented data can therefore be considered dialectally equivalent to original recordings from the perspective of linguistic content. This makes RVC a methodologically validated tool for dialect data augmentation in settings where speaker diversity is limited, with direct applicability to low-resource dialect NLP pipelines.

While the present findings are interpretable within a coherent theoretical framework, they also open specific empirical questions that the current design cannot resolve. Two directions for future research follow directly from these findings. First, the present design used a between-subjects comparison; a within-subjects replication would allow direct modeling of individual-level stability and increase sensitivity for detecting subtle condition differences. Second, the formant shifts introduced by RVC (particularly $\Delta F1 \approx 90\,\mathrm{Hz}$) raise the question of whether dialect-specific vowel contrasts are preserved at the phonological level, not merely at the perceptual level. While the perceptual data suggest that these shifts do not distort dialect-relevant contrasts, a targeted analysis of vowel space geometry before and after conversion, stratified by dialect region, could provide additional phonological corroboration of this conclusion.

5. Conclusion

This study pursued two objectives: to evaluate whether Retrieval-based Voice Conversion constitutes a valid normalization technique for dialectal speech resources, and to determine what the response of human and machine systems to this manipulation reveals about the architecture of dialect recognition. The answer is affirmative, but the more consequential finding concerns what RVC reveals about the nature of dialect recognition itself. The sharp dissociation between human stability and model improvement under identical acoustic manipulation demonstrates that these two systems do not share a common representational basis for dialect classification. Human listeners deploy speaker normalization as a cognitive default, rendering their performance robust to external speaker transformations. Neural classifiers, by contrast, are sensitive to the statistical structure of the training signal and benefit when speaker-level variance is reduced externally.

This dissociation has methodological and theoretical implications that extend beyond dialect classification. Methodologically, it establishes that human perception studies and computational modeling address different questions about speech, and that improvements in one domain do not automatically transfer to the other. Theoretically, it suggests that the gap between human and machine speech processing is not merely quantitative—which is a matter of performance levels—but qualitative, rooted in fundamentally different representational architectures. For the applied question of whether RVC constitutes a valid normalization technique for dialectal speech resources, the answer is unambiguous: dialectally relevant information is preserved, and RVC-augmented data are perceptually equivalent to original recordings. For the broader question of what RVC reveals about speech processing, the answer is more consequential. The human-machine dissociation documented here suggests that the architecture of dialect recognition—biological or artificial—determines how systems respond to speaker normalization. This architectural difference cannot be read off from performance metrics alone.

6. Limitations

The exclusive use of older male speakers limits the generalizability of the findings to mixed-gender and mixed-age settings; future work should examine whether the observed dissociation holds across speaker sex and age.

7. Acknowledgements

This research is supported by the Federal Ministry of Research, Technology and Space (BMFTR) (grant AnDy 16DKWN007) and the Academy of Sciences and Literature Mainz (grant REDE 0404), the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence LAMARR22B, and the Research Center Deutscher Sprachatlas Marburg. We are grateful to two anonymous reviewers for helpful comments.

8. Bibliographical References

Christina A. Anders. 2010. Die wahrnehmungsdialektologische Rekodierung von laienlinguistischem Alltagswissen. *Perceptual Dialectology: Neue Wege der Dialektologie*, pages 67–87.

Adam L. Anwyl-Irvine, Jessica Massonnié, Amy Flitton, Natasha Z. Kirkham, and John K. Evershed. 2020. Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods*, 52:388–407.

Matthew Baas and Herman Kamper. 2022. Voice Conversion Can Improve ASR in Very Low-Resource Settings. In *Proc. Interspeech 2022*, pages 3513–3517.

Cynthia G Clopper and David B Pisoni. 2004. Some acoustic cues for the perceptual categorization of american english regional dialects. *journal of phonetics*, 32(1):111–140.

Gunnar Fant. 1971. *Acoustic Theory of Speech Production*. De Gruyter Mouton, Berlin, Boston.

Lea Fischbach. 2024. A comparative analysis of speaker diarization models: Creating a dataset for German dialectal speech. In *Proceedings of the Third Workshop on NLP Applications to Field Linguistics*, pages 43–51, Bangkok, Thailand. Association for Computational Linguistics.

Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. 2025a. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Interspeech 2025*, pages 2780–2784.

Lea Fischbach, Akbar Karimi, Alfred Lameli, and Lucie Flek. 2025b. EDAudio: Easy data augmentation for dialectal audio. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 363–368, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Marja Gackstatter and Oliver Niebuhr. 2013. Eine kontrastive phonetische analyse niederdeutscher langvokale. *Linguistik Online*, 53(3).

Karzan Ghafoor, Sarkhel Taher, Karwan Hama Rawf, and Ayub Abdulrahman. 2025. The improved Kurdish dialect classification using data augmentation and ANOVA-based feature selection. *ARO - The Scientific Journal of Koya University*, 13:94–103.

Stephen D. Goldinger. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105(2):251–279.

Lorenz Gutscher and Michael Pucher. 2025. Audio-based classification and geographic regression of austrian dialects. In *Proc. Interspeech 2025*, pages 2765–2769.

Nina Hosseini-Kivanani, Christoph Schommer, and Peter Gilles. 2025. Voices of luxembourg: Tackling dialect diversity in a low-resource setting. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 143–152.

Markus Hundt, Nicole Palliwoda, and Saskia Schröder. 2015. Wahrnehmungsdialektologie – Der deutsche Sprachraum aus der Sicht linguistischer Laien: Exemplarische Ergebnisse des Kieler DFG-Projekts. In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen: Projekte und Perspektiven*, pages 585–629. De Gruyter Mouton, Berlin, Boston.

Keith Johnson. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson and John W. Mullennix, editors, *Talker Variability in Speech Processing*, pages 145–165. Academic Press, San Diego.

Keith Johnson and Matthias J Sjerps. 2021. Speaker normalization in speech perception. *The handbook of speech perception*, pages 145–176.

Caroline Kleen. 2022. Identifikation von regionalakzenten durch linguistische laien. Master's thesis, University of Trier. Unpublished.

Wolf Peter Klein. 2021. *Was denken linguistische Laien über die (deutsche) Grammatik?: Beobachtungen und Interpretationen anhand des öffentlichen Sprachgebrauchs*, pages 227–248. De Gruyter, Berlin, Boston.

Stefan Kleiner. 2017. F1/f2-diagramme als darstellungsmittel bairisch geprägter standardsprachlicher vokalsysteme. In Alexandra N. Lenz, Ludwig Maximilian Breuer, Tim Kallenborn, Peter Ernst, Manfred Michael Glauninger, and Franz Patocka, editors, *Bayerisch-österreichische Varietäten zu Beginn des 21. Jahrhunderts: theorie- und korpusbasierte Ansätze*, volume 167 of *Zeitschrift für Dialektologie und Linguistik – Beihefte*, pages 263–284. Franz Steiner Verlag, Stuttgart.

Olaf Köster, Roland Kehrein, Karen Masthoff, and Yasmin Hadj Boubaker. 2012. The tell-tale accent: Identification of regionally marked speech in German telephone conversations by forensic phoneticians. *International Journal of Speech Language and the Law*, 19(1):51–71.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*, volume 54. De Gruyter.

Alfred Lameli. 2025. *Gesprochenes Deutsch in den Regionen: Eine Standortbestimmung für die Bundesrepublik Deutschland*, pages 51–80. De Gruyter, Berlin, Boston.

Prolific. 2024. Prolific (version used: [May 2025]). Online platform.

Verena Sauer and Toke Hoffmeister. 2022. *Wahrnehmungsdialektologie*. De Gruyter, Berlin, Boston.

Jürgen Erich Schmidt and Joachim Herrgen and Roland Kehrein and Alfred Lameli and Hanna Fischer. 2020–. *Regionalsprache.de: Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Philipps-Universität Marburg. Digitale Sprachressource. Bearbeitet von Lisa Dücker, Robert Engsterhold, Marina Frank, Heiko Girnth, Simon Kasper, Juliane Limper, Salome Lipfert, Georg Oberdorfer, Tillmann Pistor, Anna Wolańska. Unter Mitarbeit von Dennis Beitel, Lea Fischbach, Milena Gropp, Heiko Kammers, Maria Luisa Krapp, Vanessa Lang, Salome Lipfert, Nathalie Mederake, Jeffrey Pheiff, Bernd Vielsmeier. Studentische Hilfskräfte.

W. F. Sendlmeier and J. Seebode. 2006. Formantkarten des deutschen Vokalsystems. https://www.kw.tu-berlin.de/fileadmin/a01311100/Formantkarten_des_deutschen_Vokalsystems_01.pdf.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157.

Matthias J Sjerps, Neal P Fox, Keith Johnson, and Edward F Chang. 2019. Speaker-normalized sound representations in the human auditory cortex. *Nature communications*, 10(1):2465.

Georg Wenker. 2013. *Schriften zum "Sprachatlas des Deutschen Reichs". Gesamtausgabe*. Number 111.1–3 in Deutsche Dialektgeographie. Olms, Hildesheim, New York, Zürich.

9. Language Resource References

Paul Boersma and David Weenink. 2022. *Praat: doing phonetics by computer [Computer program]*. Version 6.2.14, retrieved 24 May 2022.

Ramon Corretge. 2012–2024. *Praat Vocal Toolkit [Software extension for Praat]*. Retrieved 20 January 2024.

Yannick Jadoul and Bill Thompson and Bart de Boer. 2018. *Parselmouth: A Python interface to Praat [Software]*. Version 0.4.1, retrieved September 2022.

Joel Shor and Subhashini Venugopalan. 2022. *TRILLsson: Distilled Universal Paralinguistic Speech Representations [Pre-trained Model]*.

South Tyrolean Dialect-to-Standard Speech Translation: A Resource

Greta H. Franzini, Luca Ducceschi
Institute for Applied Linguistics, Eurac Research
Viale Druso 1, 39100 Bolzano, Italy
{greta.franzini, luca.ducceschi}@eurac.edu

Abstract

This paper presents a developing oral resource for South Tyrolean, a German dialect spoken in Northern Italy. The dialect is ubiquitous in spoken communication but lacks a standardised orthography. In this context, strict transcription into dialect is of limited to no utility to the local community. Instead, there is a distinct and strong demand for technology capable of directly translating spoken dialect into Standard German. To address this specific need, we introduce a dynamic, incrementally growing dataset designed to fine-tune ASR models for the task. Our corpus aggregates diverse sources, including media and research interviews, totalling over 13 hours of aligned audio. We describe a collaborative workflow where community partners contribute audio material in exchange for automated transcriptions, creating a virtuous cycle of data improvement. Additionally, we detail our iterative model fine-tuning strategy, the preprocessing workflow and the resulting improvements in model performance.

Keywords: dialect, German, ASR, language resource, speech corpus

1. Introduction

South Tyrolean, a Southern Bavarian variety spoken by over 90% of the German-speaking population in South Tyrol (Italy) (ASTAT, 2024, p. 1), exists in a state of *medial diglossia* (Auer, 2005, p. 12). While the dialect dominates oral communication, Standard German (Ammon et al., 2016) remains the exclusive medium for formal written domains (Leonardi, 2020). This functional split creates a distinct local demand for NLP: rather than transcriptions into written dialect —a preference mirrored in other Bavarian contexts (Blaschke et al., 2024)— users require technology that translates spoken dialect directly into Standard German. This functional separation between spoken dialect and written standard language means that South Tyrolean remains underrepresented in digitised and annotated speech resources, leaving it a low-resource language despite its high vitality.

Modelling this speech is further challenged by internal fragmentation and language contact. South Tyrol's mountainous geography has fostered distinct varieties (e.g., Puster Valley vs. Vinschgau) with divergent phonology and lexicons, meaning that models calibrated for speakers of the capital, Bozen, often fail in rural side valleys. Furthermore, frequent code-switching with Italian introduces loanwords such as *targa* ('license plate') or *magari* ('perhaps'), which often cause monolingual German models to hallucinate phonetically similar German terms. Together, all of these hurdles necessitate specialised resources for robust ASR in the region.

Contribution. While numerous oral corpora exist for German dialects, with Blaschke et al. (2023) identifying 39 audio resources documenting both Low and High German varieties, no dataset to date has been specifically designed to support dialect-to-standard ASR for South Tyrolean. We address this gap by presenting a growing oral resource developed through an increasingly collaborative, community-driven data collection approach, together with a family of fine-tuned ASR models capable of translating South Tyrolean dialectal speech directly into Standard German (Ducceschi and Franzini, 2025).

To facilitate the interpretation of the qualitative error analysis presented later in the paper, it is worth noting that translation from South Tyrolean dialect into Standard German is not a purely literal mapping task. Differences between spoken dialectal usage and written standard language frequently require structural rephrasing rather than word-for-word rendering. As a result, some model outputs may appear acoustically plausible while nonetheless diverging from the expected Standard German form in terms of morphology, syntax or lexical choice. This distinction is particularly relevant for understanding the error patterns discussed in Section 4.2.

2. Related Work

The development of ASR systems for low-resource languages has gained traction in recent years, yet dialectal varieties remain under-served. The core challenge lies in the data bottleneck: state-of-the-art models like Whisper require thousands of hours of transcribed audio to generalise effectively. While Standard German is supported by vast corpora, its dialectal offshoots often have negligible digital

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 188–194
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

footprints.

Previous work in Germanic dialect-to-standard ASR has concentrated significantly on Swiss German. Initiatives in Switzerland have produced aligned corpora, such as *STT4SG-350* (Plüss et al., 2023), *PASSAGE* (Gerlach et al., 2022), *GermEval 2020 Task 4* (Plüss et al., 2020), *SRF Meteo* (Stadtschnitzer and Schmidt, 2018) or *ArchiMob* (Scherrer et al., 2019). While *ArchiMob* primarily employs a word-aligned normalisation layer that provides Standard German equivalents at the lexical level, these resources collectively facilitate the development of robust models for dialect-to-standard translation. Similar developmental efforts have also been documented for Luxembourgish (Gilles et al., 2023), a related West Germanic language, and Bavarian (Blaschke et al., 2025).

3. Method

3.1. Data Collection and Composition

Our dataset is a dynamic, work-in-progress resource that is incrementally extended as new data becomes available. As Table 1 shows, we aggregate data from a diverse set of sources to ensure coverage of different domains, geographical areas and speech styles.

Crowdsourced Corpus (ALP). We incorporate the South Tyrolean German subset of the AlpiLinK (Alpine Languages in Contact) corpus (Kruijt and Rabanus, 2025). This openly available resource provides crowdsourced oral data across the Alpine regions of Italy (Rabanus et al., 2025). We specifically use the translation task, which features a controlled set of 30 Standard German written sentences orally rendered into dialect by native speakers. This structured repetition provides the model with multiple phonetic realisations of identical lexical content, aiding the mapping between dialectal variation and the Standard German reference.

Subtitled YouTube Videos (IDM). We harvest publicly available video content that already possesses creator-provided subtitles. Specifically, we collected spontaneous speech data from 44 subtitled promotional videos published by *IDM Südtirol*[1]. These videos feature unscripted interviews with speakers of varying ages (excluding children), covering diverse topics from cooking to farming. The videos were downloaded as `MP4` files using the Python *pytube* library[2]. To extract the subtitles, we processed each video frame by enhancing contrast and applying *PaddleOCR* (Kanakaraddi et al., 2024) to the bottom third of the image. OCR errors were primarily due to special character misrecognition (e.g., ß, ä, ü, ö), low video resolution and visual interference. All output was manually corrected to produce clean reference texts. Since the original subtitles are often paraphrased rather than being verbatim transcriptions, the alignment between audio and text is approximate rather than literal.

Language Learning Textbooks (ARB). To ground the model in more structured speech, we use audio materials from dialect teaching resources, which are typically clearer and more grammatically standard than the YouTube data. We used scripted speech data from the publicly available *DaZUgeHÖREN* textbook, designed for learners of South Tyrolean dialect (Gurschler and Tscholl, 2015). The textbook includes audio exercises on CD, featuring dialectal prompts accompanied by printed Standard German translations. Again, the alignment between audio and text is not always exact, as the printed translations occasionally deviate from the spoken content.

Audiovisual Archives (FUM, MEN). We collaborate with cultural institutions and historical research projects to access archival data. As part of a formal agreement with the Office for Film and Media of the Province of Bolzano (FUM), we are transcribing their audiovisual archives, primarily 30+ year old news broadcasts containing both South Tyrolean Standard German and dialect segments, using our best-performing ASR model. Additionally, we have an informal agreement with an Austria-based publishing agency (MEN) producing documentary films on South Tyrolean history. These documentaries feature elderly native speakers with broad, linguistically challenging dialect forms.

Research and Purpose-recorded Interviews (EUR). Beyond external partnerships, we assist other colleagues within our organisation by transcribing their research interviews, thereby incorporating their dialect data into our corpus. Furthermore, we address data gaps through targeted elicitation tasks. We record human translators spontaneously producing spoken dialect translations of Standard German sentences, following a methodology similar to Plüss et al. (2022, p. 2).[3] To enhance the model's performance on local named entities, a frequent stumbling block, we use our company instance of Microsoft Copilot to generate Standard German prompts rich in local terms for these recording sessions.

All audio data is segmented into chunks shorter than 30 seconds, a requirement for effective training with architectures such as Whisper. The files are converted to `WAV` format and resampled to 16 kHz using *ffmpeg* (FFmpeg Developers, 2016), then manually aligned with their respective tran-

[1] https://www.youtube.com/@suedtirol.official

[2] https://pytube.io/en/latest/index.html

[3] Recordings are made in segments of under 30 seconds using a dynamic cardioid Behringer ULTRAVOICE XM8500 vocal microphone.

scriptions in ELAN[4]. A custom Python script subsequently parses the resulting ELAN alignments (EAF) and compiles the final dataset into a single aggregated CSV file for model fine-tuning.

3.2. Collaborative Workflow

Our methodology increasingly centers on a mutually beneficial partnership with internal and external research and memory institutions. Partners provide raw audio archives in exchange for automated Standard German transcripts generated by our latest model. With typical accuracies of 70–80%, partners only correct errors in ELAN rather than transcribing from scratch, significantly reducing manual effort (Russell et al., 2024). Once manually verified, this high-quality data is integrated into the training set for the next fine-tuning cycle. This creates a virtuous cycle where continuous model improvement further decreases partner workload over time.

3.3. Preprocessing and Quality Control

Given that noise or inconsistencies introduced prior to translation may negatively affect downstream performance, we explicitly assessed potential sources of error in the preprocessing pipeline. As mentioned in Section 3.1, all audio was used as retrieved from the original providers and underwent only minimal, standardised processing: conversion to WAV format and resampling to 16 kHz. No signal enhancement, denoising, filtering or speaker normalisation was applied, as our goal was to evaluate translation performance under realistic data conditions rather than optimise acoustic quality.

To avoid segmentation-related artefacts, audio files were manually segmented in ELAN into units shorter than 30 seconds, with segment boundaries placed in silence regions to prevent truncation of words or utterances. Where written translations were missing or derived from noisy sources (e.g. OCR-extracted subtitles), these were —as mentioned in sections 3.1 and 3.2— produced or corrected and subsequently validated. Aside from differences in translation availability and degree of audio–text literalness across sources, the preprocessing pipeline was uniform across all data types.

To characterise the acoustic conditions of the dataset, we assessed its technical quality using *librosa* (McFee et al., 2015). Overall, the corpus exhibits high acoustic quality, with a mean Signal-to-Noise Ratio (SNR) of 40.35 dB and no instances of significant digital clipping ($>$1%). A small subset of the data (26 files from ALP) shows poor SNR ($<$10 dB), reflecting the heterogeneous and sometimes noisy recording conditions typical of real-world material. In addition to this quantitative assessment, we performed qualitative preprocessing checks focussing on segmentation consistency and gross audio integrity. While we did not conduct a systematic or quantitative analysis of audio–text correspondence —since this lies outside the scope of the present work— these checks did not reveal pervasive segmentation errors, severe signal degradation or obvious audio–text inconsistencies that would plausibly account for the translation errors observed in evaluation. Residual variability in recording conditions is therefore treated as an inherent property of the dataset rather than as a preprocessing artefact.

3.4. Dataset Evolution and Tuning Strategy

We employ a rigorous iterative process for dataset construction and model tuning. With every new data increment, the model is fine-tuned to enhance its performance and generalisation capabilities. To ensure the integrity of our resources, we enforce strict consistency checks: each subsequent version of the dataset is verified to be a strict superset of its predecessor.

We maintain an 80/20 train-test split for each version, allowing us to track the model's performance on the specific distribution of each data increment. At the same time, to ensure longitudinal comparability and track performance improvements that are strictly attributable to training data scale rather than changes in test set difficulty, we evaluate all model versions against a common benchmark: the latest test set, v1.3. Currently, the resource comprises over 10 hours of training data and nearly 3 hours of test data.

4. Results

The growth of the dataset and the resulting ASR performance are shown in Table 2. We fine-tuned a separate OpenAI Whisper large-v3 model for each dataset version. We then evaluated all fine-tuned models on the v1.3 test set (1,340 samples, 24,442 tokens, 2h 50m) using both WER and BLEU. While BLEU is typically most robust when calculated against multiple reference translations, we include it here to complement WER in capturing translation quality. This is particularly relevant as our Standard German references occasionally exhibit significant structural or lexical deviations from the original dialectal utterances, moving beyond simple word-for-word mappings. To ensure the metrics reflect semantic and lexical accuracy rather than orthographic or formatting differences, we normalise both the hypothesis and reference texts by

[4]Version 6.4. https://archive.mpi.nl/tla/elan

Table 1: Current overview of corpus (v1.3). The entire audio corpus was segmented into 30-second units, each manually paired with the corresponding written translation.

Source	Type	Speech	Hours	Speakers	Configuration	Age
ARB	transcribed audio	read	47m	3	single-speaker	20-49
ALP	transcribed audio	read, spontaneous	4h 47m	180	single-speaker	10-89
EUR	audio	read, spontaneous	3h 42m	3	multi-speaker	20-49
FUM	video	read, spontaneous	18m	7	multi-speaker	0-79
IDM	subtitled video	spontaneous	1h 03m	86	multi-speaker	20-79
MEN	video	spontaneous	1h 35m	2	multi-speaker	30-99
TOTAL			13h 16m			

Table 2: Training data evolution and ASR performance. All models, including the non-fine-tuned OpenAI Whisper large-v3, are evaluated on the v1.3 test set (1,340 samples, 24,442 tokens, 2h 50m), with v1.2 yielding the best overall performance.

	Training Data				Performance (v1.3)	
Model	**Samples**	**Tokens**	**Avg.L(s)**	**Duration**	**WER ↓**	**BLEU ↑**
OpenAI Whisper large-v3		*n/a (no fine-tuning)*			0.46	44.58
v1.0	4,441	51,474	9.79	6h 39m	0.37	0.52
v1.1	5,048	75,203	11.46	9h 07m	0.27	65.65
v1.2	5,325	87,298	12.26	10h 16m	**0.24**	**69.13**
v1.3	5,363	88,810	12.28	10h 26m	0.24	68.73

lowercasing and removing punctuation prior to evaluation, thus mitigating WER inflation.

4.1. Quantitative Performance

The results shown in Table 2 showcase a clear trajectory of improvement. Notably, already with v1.0 of the dataset —despite its small size— our fine-tuned model obtains decent results. This aligns with the findings of Hollenstein and Aepli (Hollenstein and Aepli, 2014) as well as Samardžić et al. (Samardžić et al., 2015) for Swiss German, who demonstrated that modest amounts of variety-specific data often outperform significantly larger, out-of-domain Standard German corpora in training language processing tools.

The transition from v1.0 to v1.1, which increased training data from roughly 6.5 to 9 hours, yielded the single largest performance gain, reducing WER from 0.37 to 0.27 (a 27% relative improvement). This jump coincides with the introduction of archival interviews and additional purpose-recorded data, which substantially diversified the speaker pool and exposed the model to more varied spontaneous speech. Performance continues to improve with v1.2 (WER 0.24), as broadcast data and further subtitled content bring additional domain coverage. Between v1.2 and v1.3, global WER remains stable at 0.24 while BLEU decreases marginally from 69.13 to 68.73.[5] This plateau is expected: v1.3 adds only 38 new training samples over v1.2, contributing less than 10 additional minutes of audio. Rather than a performance regression, the near-identical scores confirm that the model has effectively saturated on the available data, consistent with the logarithmic relationship between training data size and ASR accuracy observed in large-scale studies (Zhang et al., 2023; Radford et al., 2023). This signals that further gains will require substantive new data increments rather than incremental additions.

4.2. Qualitative Analysis

A qualitative error analysis reveals that the model often confuses inflectional variants. In several instances, it outputs common indicative or finite forms, such as *war* ('was'), *gab* ('there was/were') and *sind* ('are'), instead of the grammatically required subjunctive or infinitive forms, that is *wäre* ('would be'), *gäbe* ('there would be') and *sein* ('to be'). While such distinctions are linguistically subtle, their repeated occurrence suggests systematic difficulties in distinguishing between closely related verbal paradigms. More broadly, the model sometimes fails to disambiguate lexical items that exist in both Standard German and South Tyrolean dialect but differ in meaning. In such cases, it tends to reproduce the acoustically plausible form even when contextually inappropriate. For example, the dialectal form *wert*, intended as a realisation of *wird* ('will'), was transcribed as *wert* ('worth'), although this interpretation rendered the utterance semanti-

[5] BLEU scores are computed using SacreBLEU with signature nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1

cally incoherent in context.

Furthermore, characteristic South Tyrolean forms such as *sem* ('there/at that time'), *homo* ('we have'), *hon* ('I have'), *kimmp* ('comes') and *man* ('mine') are not always rendered accurately, but are generally processed satisfactorily.

The multilingual environment of South Tyrol naturally influences model behaviour. Given the dialect-heavy and translation-oriented training setup, foreign lexical items may be retained, translated or inconsistently transliterated. In one case, the model correctly translated the Italian word *aziendale* ('corporate') as *Unternehmerisch* while in another it incorrectly transliterated the same word as *achtsindale*.

Automated punctuation continues to pose difficulties. The model often struggles to correctly identify sentence boundaries and to insert commas where appropriate. Speaker attribution is likewise challenging. However, this limitation appears to stem primarily from the diarisation component employed, specifically pyannote.audio*pyannote.audio* (Bredin et al., 2019), rather than from the speech recognition model itself.

Finally, the model demonstrates limitations in the recognition of named entities. Place names and personal names are frequently misidentified or inconsistently rendered (e.g., *Potsdn* instead of *Bozen*; *Aldomoro* instead of *Aldo Moro*).

5. Discussion

The collaborative framework successfully addresses the data bottleneck by transforming the reuse and enhancement of audiovisual and research recordings into a driver for technological democratisation. A key finding is the model's ability to maintain semantic coherence across spontaneous, non-linear dialectal utterances, effectively mapping fragmented speech into structured Standard German prose. This functional utility is particularly evident for our partners, for whom the system provides a reliable baseline that significantly accelerates the transcription process, despite the need for human correction.

The results indicate that the model remains robust despite the presence of a subset of messy real-world recordings, as evidenced by the high overall performance despite the 26 files with poor SNR, although certain linguistic nuances remain a challenge. The persistent issues with named entities, as well as specific lexical items and verb forms suggest that future work should focus on data augmentation in these areas.

During the manual revision process, we document these acoustic and linguistic challenges to inform our filtering strategies and provide feedback to partners. This project thus serves as both a technical resource and a capacity-building initiative for the local speaker community.

6. Conclusion and Future Work

This project responds directly to community demands for accessible communication between South Tyrolean dialect and Standard German. By implementing a community-oriented framework, we have established the first dedicated dataset for the region. Our virtuous cycle methodology demonstrates that partnering with institutions can effectively turn various materials into high-quality training data, resulting in a system that is already perceived by users as a practically valuable tool for transcription and documentation.

The resource, however, remains a work in progress. Future work will prioritise filling geographic gaps to ensure the model generalises across South Tyrol's dialectal landscape. We also plan to experiment with larger model architectures, such as NVIDIA Canary, and evaluate whether the integration of Austrian Standard German (e.g., from the *Graz corpus of read and spontaneous speech* (Schuppler et al., 2014)) can further improve robustness.

7. Data Availability

The versioned resource repository is publicly available.[6] While the repository provides comprehensive metadata for all data included in the resource, licensing constraints prevent us from redistributing a substantial portion of the current corpus.

8. Limitations

A primary limitation of this study is that the corpus does not yet provide exhaustive coverage of all South Tyrolean dialectal variants, with certain valleys remaining underrepresented in the current training data. Additionally, while SNR provides a measure of background noise levels, it does not fully capture the acoustic complexity of our corpus. Future iterations will use non-intrusive neural metrics like NISQA (Mittag et al., 2021) to better model the relationship between acoustic quality and translation accuracy.

9. Ethical Considerations

Participants of the EUR corpus consent to the use of their voice and their involvement is entirely voluntary. No metadata or personal identifiable infor-

[6] https://gitlab.inf.unibz.it/commul/speech-to-text/augusta_data/

mation is used in the training or evaluation of the model.

10. Acknowledgements

We extend our sincere thanks to all project partners for their contributions to this work. We are also grateful to the anonymous reviewers for their valuable comments and suggestions, and to Simone Baratella for his technical support in the acoustic assessment of the training data.

11. Bibliographical References

Ulrich Ammon, Hans Bickel, and Alexandra N. Lenz, editors. 2016. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*, 2nd edition. De Gruyter, Berlin.

ASTAT. 2024. Sprachkenntnisse und Sprachgebrauch im Alltag in Südtirol - 2024.

Peter Auer. 2005. *Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations*, pages 7–42. De Gruyter Mouton, Berlin, New York.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A Survey of Corpora for Germanic Low-Resource Languages and Dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, page 913–917. ISCA.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. pyannote.audio: neural building blocks for speaker diarization.

Luca Ducceschi and Greta H. Franzini. 2025. Speech transcription from South Tyrolean Dialect to Standard German with Whisper. In *Interspeech 2025*, page 5, Rotterdam, The Netherlands.

Ludwig M. Eichinger. 2002. South Tyrol: German and Italian in a Changing World. *Journal of Multilingual and Multicultural Development*, 23(1-2):137–149.

FFmpeg Developers. 2016. ffmpeg tool.

Johanna Gerlach, Jonathan Mutal, and Pierrette Bouillon. 2022. Producing Standard German subtitles for Swiss German TV content. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 37–43, Dublin, Ireland. Association for Computational Linguistics.

Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023. ASRLUX: Automatic Speech Recognition for the low-resource language Luxembourgish. In *Proceedings of the 20th International Congress of Phonetic Sciences*. Guarant International.

Michael Gurschler and Evi Rita Tscholl. 2015. *DaZUgeHÖREN: Südtiroler Dialekt von Jugendlichen für Jugendliche Arbeitsmaterialien zum Südtiroler Dialekt*. Autonome Provinz Bozen - Deutsches Bildungsressort.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Suvarna Kanakaraddi, Ashok Chikaraddi, Shantala Giraddi, Karuna Gull, and Mallanagouda Patil. 2024. Optimized Scene Text Detector and Paddle Optical Character Recognizer Techniques to Extract Text from Images. In *Proceedings of 4th International Conference on Artificial Intelligence and Smart Energy*, pages 218–230, Cham. Springer Nature Switzerland.

Anne Kruijt and Stefan Rabanus. 2025. From VinKo to AlpiLinK: web-based long-term storage and accessibility of information. *Korpus im Text*, 17.

Mara Maya Victoria Leonardi. 2020. "I hardly ever practice the real Standard German." Self-reported language use and language proficiency in South Tyrol (Italy). *Linguistik Online*, 102(2):83–98.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy 2015*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, page 2127–2131. ISCA.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German Speech to Standard German Text Corpus.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. GermEval 2020 Task 4: Low-Resource Speech-to-Text. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland. CEUR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sam O'Connor Russell, Iona Gessinger, Anna Krason, Gabriella Vigliocco, and Naomi Harte. 2024. What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics*, 3(3):100163.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 4th biennial workshop on less-resourced languages*, pages 294–298. ELRA.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454.

Barbara Schuppler, Martin Hagmüller, Juan Andres Morales-Cordovilla, and Hannes Pessentheiner. 2014. GRASS: the Graz corpus of Read And Spontaneous Speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1465–1470.

Michael Stadtschnitzer and Christoph Schmidt. 2018. Data-Driven Pronunciation Modeling of Swiss German Dialectal Speech for Automatic Speech Recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages.

12. Language Resource References

Rabanus, Stefan and Kruijt, Anne and Alber, Birgit and Bidese, Ermenegildo and Gaeta, Livio and Raimondi, Gianmario. 2025. *AlpiLinK Corpus*. University of Verona and Free University of Bolzano, 1.2.0. PID 10.5281/zenodo.8360169.

TransVar – the Corpus for Variation and Change Study of the Historical Transcarpathian lects

Ilia Afanasev
University of Vienna
Vienna, Austria
ilia.afanasev.1997@gmail.com

Abstract

The paper introduces TransVar – the corpus of the historical Transcarpathian lects (the first half of the XX century, the territories of modern Ukraine, Poland, Slovakia, and Romania). The corpus contains data from Lemko, Bojko and Hutsul small territorial lect groups. It is crucial for studies of the people of these territories, who witnessed forceful deportation from their homeland in the 1940s – 1950s, soon after the recordings were made (1920s – 1930s). The article also provides a brief overview of their linguistic properties, as evident in the material.

The corpus is morphosyntactically tagged. It contains data on part-of-speech, morphological features, lemmata and syntactical dependencies. The study stresses the crux of manual analysis of the errata made in an automatic tagging phase for further improvement. The supplementary information includes named entities encountered in the text and the basic vocabulary. All the texts are accompanied by metalinguistic information, required for the sociolinguistic study.

After the analysis of the current stage of the corpus creation, the article outlines further research prospects. Apart from more thorough manual annotation, one of the prospects is to add English translation with the purpose of making the material more accessible to scholars without a background in Slavic studies.

Keywords: Transcarpathian, sociolinguistics, Universal Dependencies, low-resourced language modelling, part-of-speech tagging, lemmatisation, dependency parsing, qualitative analysis

1. Introduction

Historical low-resourced, small territorial lects[1] are probably among the most understudied types of resources in computational linguistics (Miletić and Siewert, 2023). This material is severely low-resourced and highly non-standard, lacking normalisation. This makes it especially challenging for NLP tools, even in comparison with modern non-standard varieties (Piotrowski, 2012). Still, there is a significant number of resources containing texts produced by speakers of these lects. This article considers material gathered in the Transcarpathian region in the twentieth century (in the territories of modern Ukraine, Poland, Slovakia, and Romania) by several groups of scholars and published either concurrently (Nakonečna and Rudnyc'kyj, 1940) or subsequently (Kuraszkiewicz, 1963; Zheguc, 2001), and presents a corpus based on this material. An overview of the resources used to compile the corpus is provided in Section 4.

All the lects are East Slavic[2]. They belong to three main groups: Bojko, Lemko, and Hutsul (Section 3 describes their linguistic features and position within the Slavic clade). These lects have never been the subject of a quantitative variationist study due to the absence of an open-access corpus—a gap this study intends to close (Section 2 contains a short overview of variationist studies of East Slavic languages, including the Transcarpathian lects, using NLP tools).

The corpus contains part-of-speech and morphological tags, along with lemmatisation and dependency parsing. It also includes specific tags for basic vocabulary items and named entities that facilitate historical research. Each text is accompanied by a set of extralinguistic (metadata) tags containing information about the speaker, the dialectologist who conducted the recording, and the source itself. Section 5 outlines the corpus design, the tagging schema, and the annotation tools. It

[1] The article denotes any given language variety (idiolect, doculect, dialect, regiolect, sociolect, standard) as *lect*. This choice enables the elimination of overly hierarchical dichotomies, namely, `X is a dialect of Y`, which imply some degree of **X** being a lesser version of **Y**; the study of **X** can thus only be comparative, guided by differences from **Y**. Rejecting this differential approach is paramount for building a corpus, an integrative endeavour by definition (Goldin and Kryuchkova, 2011).

[2] Within historical Slavic studies, there is no consensus on whether West Slavic, South Slavic, and East Slavic are valid groupings or whether Proto-West Slavic, Proto-South Slavic, or Proto-East Slavic actually existed (Nikolaev, 1988, p. 116; Krys'ko, 1998, p.85–89; Zaliznyak, 2004, p. 7–8; Matasović, 2008, p. 83; Saenko, 2020). This article uses the term *East Slavic* in a purely geographical sense to denote the eastern part of the Slavic continuum, including the territory from eastern Slovakia in the south-west to the coast of the White Sea in the north-east.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 195–208
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

also discusses the importance of shared lexicons for NLP in low-resourced settings. Section 6 places this analysis in the context of possible future improvements.

1.1. Contribution

The study introduces a new resource for the study of historical Transcarpathian lects from both quantitative and qualitative perspectives: a morphosyntactically tagged corpus in the Universal Dependencies (UD) format (de Marneffe et al., 2021). It also provides a linguistic analysis of the performance of the Stanza model trained on modern Ukrainian and initiates the development of guidelines for the automatic annotation of small territorial East Slavic lects.

2. Related Work

2.1. East Slavic Small Territorial Lects Corpora

While gathering the material of small territorial lects has been a pivotal part of dialectology studies from their wake (Wenker et al., 1889–1923) for the recent two hundred years (Kalnyn', 1973; Nazarova, 1977), the push for digitisation is relatively recent, especially for the digitisation of texts (Goldin, 1990). Most digitisation efforts still consider atlases (Trüb, 1989) that get transformed into digital maps (Marchenko et al., 2025). Nevertheless, the Russian National Corpus (Kachinskaya and Sichinava, 2015) and the Belarusian N-Corpus both (National Corpus of the Belarusian Language, 2018) have dialectological subcorpora. While the General Regionally Annotated Corpus of the Ukrainian Language includes some regional variation, there are almost no autochthonous small territorial lects within the material (Shvedova et al., 2017–2025). Aside from the large national-level corpora, there are other initiatives. The most notable are the TrimCo corpus (Wiemer and Seržant, 2020), containing small territorial lects of East Baltic, including multiple Belarusian and Russian small territorial lects, and the HSE collection of the East Slavic small territorial lect corpora (Daniel et al., 2013–2018; Garder et al., 2018; Ter-Avanesova et al., 2018, 2019; Ronko et al., 2019; Ryko and Spiricheva, 2020), each containing some hundred thousand tokens, enabling medium-scale variation research, including quantitative approaches. Aside from them, there are local initiatives dedicated to thorough investigation of a particular group of lects. The Saratov dialectological corpus comprises material from three small territorial lects, Northern Russian Megra, Middle Russian Belogornoje, and Middle Russian Zemlianyje Hutora, gathered over the last forty years (Goldin, 1990; Goldin and Kryuchkova, 2011). The Tomsk dialect corpus is a digitisation of materials from many different lects of the Tomsk Region and Western Siberia in general, adding up to 3.5 million tokens (Zemicheva et al., 2023).

Overall, there is a significant number of corpora for modern East Slavic territorial lects. Yet, historical variations are clearly underrepresented, and there are no corpora for the western part of the continuum representing the small territorial lects and not the more modern regiolects (for the whole territory of the standard Ukrainian distribution and further to the West into the territories of Poland and Slovakia).

2.2. East Slavic Language Variation and NLP Tools

There is a rich tradition of variationist studies of Slavic languages in general, and East Slavic in particular. There are works dedicated to the variation in phonetics (Moroz, 2024), morphology (Ryko, 2024), lexicon (Zemicheva, 2020), and syntax (Moroz, 2016). Some take a wider typological approach (Wiemer et al., 2017), others restrain the comparison to the smaller areas (Ryko and Spiricheva, 2022).

However, there is one gap in this body of research. Despite the widening application of computational methods in variationist studies of East Slavic languages (Koile and Moroz, 2024), there is a relatively small number of studies that consider the role of variation in building language resources, specifically oriented towards East Slavic in general and Transcarpathian in particular (Scherrer and Rabus, 2017; Rabus and Scherrer, 2017), when contrasted to the other Slavic groups (Ondrejová and Šuppa, 2024; Lendvai et al., 2025). This article aims to outline the project of a corpus that facilitates closing the gap.

3. Transcarpathian Lects

This section gives a short overview of all of the Transcarpathian lects used to build the corpus from both the diachronic and synchronic perspectives.

3.1. The General Description History of the Clade

The *Transcarpathian* lects are a group of East Slavic small territorial lects spoken in the geographical region of Transcarpathia. There are three main groups, for which the researchers collected the databases: Bojko, Lemko and Hutsul (Kuraszkiewicz, 1963, pp. 67–72; Zilyns'kyj, 1933, pp. 8–10; Del Gaudio, 2017, pp. 64–85). The

first two are Slavic lects that are the closest to being autochthonous (by the terminology of Barannikova (2005, p. 193)) on these territories. The Hutsul lects are late settlement (by the terminology of Barannikova (2005, p. 193)). The map 1 from (Zilyns'kyj, 1933) shows the geographic distribution of the lects in the 1920s – 1930s.

3.2. Morphology

The scholars (Kuraszkiewicz, 1963; Myholynec', 2004; Del Gaudio, 2017) report on a set of common Transcarpathian features that divide this area from the neighbouring East Slavic languages. The most prominent features include:

- Transcarpathian lects retain short forms of personal pronouns, for instance, ç'a 'self-ACC.SG' that in some cases can undergo root reduplication: seç'a 'self-ACC.SG'. The frozen form ç'a 'self-ACC.SG' denoting the reflexivity of the verb is not restricted in its distribution within the clause, in contrast (Kuraszkiewicz, 1963, p.71).
- In Transcarpathian lects INS.PL of names has -ma, a former INS.DU ending, cf. ps'oma 'dog-INS.PL'. The standard Ukrainian has form ps'ami 'id.'.
- Present tense third person singular and plural, along with imperative, have -t ending, in contrast to standard Ukrainian ∅ or -t^{j}: Transcarpathian xod'i-t 'walk-PRES.3SG' – standard Ukrainian xod'i-t^{j} 'id.' (Myholynec', 2004, p. 18).

To summarise, there are many common archaisms within the Transcarpathian area, which alternate the texts enough for the NLP tools to struggle, presenting a well-suited material for variation study research.

3.3. Lexicon

The lexicon of Transcarpathian lects is both highly innovative and highly archaic. There are many borrowings from the neighbouring languages across all the Transcarpathian lects (Nakonečna and Rudnyc'kyj, 1940, p. 71). At the same time, there are a lot of archaisms (Del Gaudio, 2017, p. 85). One of the most prominent features is the word for 'one'. In the Transcarpathian lects it is jed'en, a form closer to West and South Slavic, while in standard it is the usual East Slavic od'in (Kuraszkiewicz, 1963, p.72). This may cause some issues, similar to the out-of-domain problem (Afanasev and Lyashevskaya, 2023), but as the Transcarpathian lects contributed a lot to the standard Ukrainian vocabulary (Del Gaudio, 2017, p. 85), not crucial ones.

3.4. Syntax

There are several peculiar constructions in the Transcarpathian lects not present in standard Ukrainian, including:

- The Transcarpathian lects use prepositions na 'on' and k 'to' to express the direction instead of the standard Ukrainian do 'towards'.
- The Transcarpathian lects use conjunction ﬁi in comparative constructions.

Some of these, like the accusative possessed construction, are less likely to cause issues with syntactic tagging. The others, like the causative construction, seem to present a bigger problem.

4. Resources

The texts that comprise the corpus are mostly a product of the collective effort of a single group of scholars, who, during the 1920 – 1930s, documented various East Slavic lects spread mostly in the Transcarpathian geographical region from Central Slovakia to Western Ukraine and from Southern Poland to Northern Romania. More texts are available via the efforts of the 1950s researchers, but these are relatively more scarce. The following paragraphs describe each of the books that came out of the efforts in terms of the presented texts, their description, and metalinguistic information.

4.1. Nakonečna and Rudnyc'kyj (1940)

Nakonečna and Rudnyc'kyj (1940) contains texts, a dictionary and general information on three southwesternmost East Slavic small territorial lects of Transcarpathian territories. The book consists of three parts: the general introduction, the description of the material lect by lect, and the short differential[3] dictionary of the lexical items from all the lects. Of these three, all are relevant for the digitisation effort and corpus creation.

The introduction starts with the general information on the linguistic features of the analyzed lects, which are common for all of the analysed lects, when they are compared to other East Slavic lects and especially Ukrainian (Nakonečna and Rudnyc'kyj, 1940, pp. 7–8).

The introduction continues by providing metainformation on the speakers of all three lects, mainly stating their linguistic and social background (date of birth, education, exposure to standard Ukrainian).

[3]The one that contains only the items that differ from the standard language, in this case, Ukrainian, as opposed to the full, which contains all the lexical items from the studied lect.

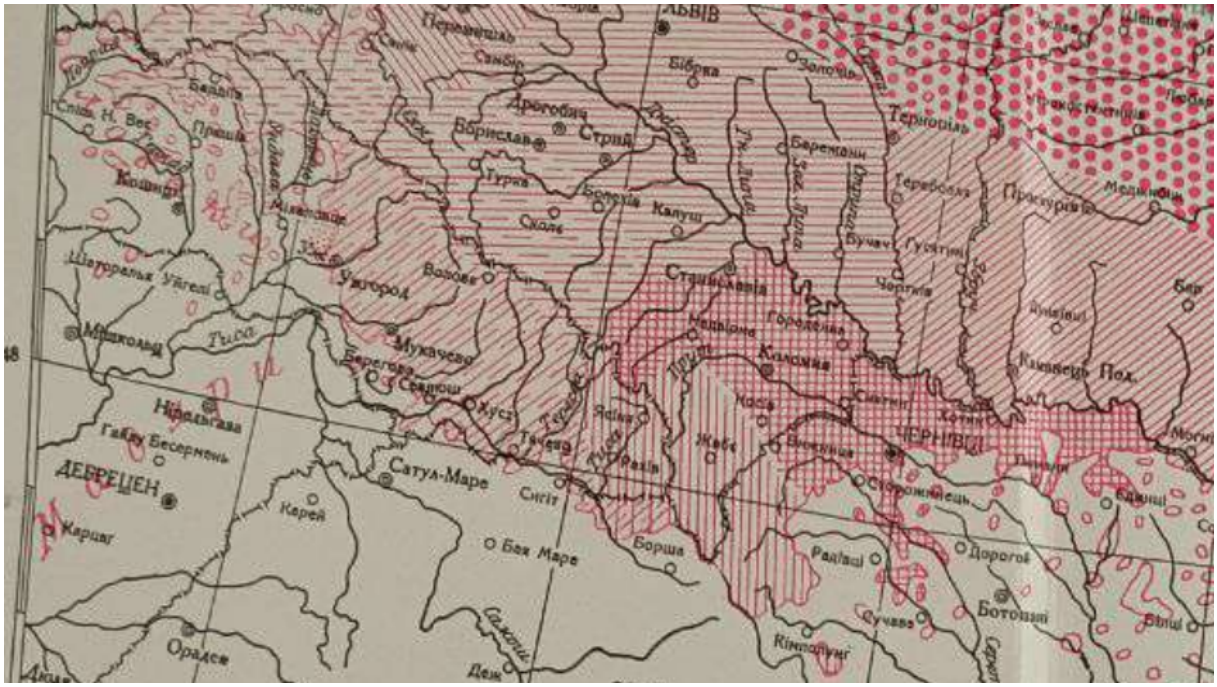


Figure 1: The map of Transcarpathian lects distribution in the beginning of the XX century (Zilyns'kyj, 1933). Rare horizontal strikes denote Lemko (in the left corner), more dense horizontal strikes (closer to the center) denote Bojko, dense vertical strikes (closer to the right) denote Hutsul.

The introduction also outlines the research methodology. The crucial part is an extremely detailed table of phonetic transcription (Nakonečna and Rudnyc'kyj, 1940, pp. 17–18) providing key information for digitising the texts in a more common format.

The next section of Nakonečna and Rudnyc'kyj (1940) consists of three parts, each dedicated to one particular idiolect from Lemko (Nakonečna and Rudnyc'kyj, 1940, pp.23–37), Bojko (Nakonečna and Rudnyc'kyj, 1940, pp.49–62) and Hutsul (Nakonečna and Rudnyc'kyj, 1940, pp. 65–82) parts of the continuum. Each of these parts outlines the phonetic, morphological, and lexical peculiarities of the lect, as well as provides the texts in these lect. The main issue is that the comparative research does not represent small territorial groups of Lemko, Bojko and Hutsul, but a single speaker out of each of these groups, which can complicate a variation study (including a historical phonology one). Metainformation on the texts (there is no split between the texts, only between – it seems – the recordings) and the speakers is not present in these parts, which is especially problematic, given that it is already significantly restricted in the introductory part of the book.

The representation of texts is rather detailed, with three forms: the phonetic transcription[4], the standardised[5] transcription, and the German translation[6]. Table 1 shows an example.

to sut fʂ'ɯtkɣ lemk'ӧwskɣ s'elɑ
de po lemk'ӧwskɣ fiw'ɑrjɑt
То сут вши́тки лемки́вски се́ла,
де по лемки́вски гва́рят
Das sind alles lemkische Dörfer,
wo lemkisch gesprochen wird.

Table 1: LA1407.1.4, an example of a sentence from Nakonečna and Rudnyc'kyj (1940, p. 31). At the top, there is phonetic representation and in the middle is a standard-based transcription. The German translation is at the bottom. The translation of the sentence is *These all are Lemko villages, where one speaks Lemko.*

The fine-grained transcription is, on the one hand, a serious advantage for the reconstruction of the phonetic system. However, there is no understanding of how reliable it actually is, especially without surviving recordings. One more issue is that there is almost no possibility to use either original transcription or its IPA rendering for automatic tagging. For the latter purposes, the study is going to use a standard-based transcription, also provided in Nakonečna and Rudnyc'kyj (1940).

After a short conclusion (Nakonečna and Rudnyc'kyj, 1940, pp. 83–85), the book provides a dic-

[4] I provide the IPA version, as it is more well-known and easy to use than the original transcription.

[5] Using the same set of graphemes as standard Ukrainian of the time.

[6] German seems to be L2 for authors, as previous research has claimed (Afanasev, 2025).

tionary of the words, specific for the small territorial lects under study. While not providing the full information required for a lexical study, this is especially helpful for lemmatisation. Overall, Nakonečna and Rudnyc'kyj (1940) provides a material of high quality for the time given, though lacking clarity and transparency.

4.2. Zheguc (2001)

Zheguc (2001) is a collection of texts, collected in different time periods (1920s – 1930s and 1990s) in the Hutsul part of the Transcarpathian region. For chronological uniformity, the corpus takes only the texts that come from the 1920s – 1930s. While this restricts the overall quantity of the material and does not allow for a proper language change study, it greatly assists in balancing the corpus. The modern material is the result of continuous field trips; thus, it is going to reduce the historical material to residuals if put in the same corpus. This is not to mention that due to the Soviet era deportations the linguistic landscape has probably drastically changed between the 1920s – 1930s and the 1990s, therefore it is hardly suitable for a proper comparison.

The transcription system of texts is clear. It is standardised phonetic/phonematic transcription, based on standard Ukrainian, but depicting the key phonetic and lexical features of the Transcarpathian lects, relevant for the purposes of the study. It does not represent all the phonetic nuances and interidiolect variation, but preserves the crucial characteristics and acts as a compromise solution, necessary for the texts for which there is no recording available. The established rules of the transcription system allow it to be unified with the standardized transcription system of Nakonečna and Rudnyc'kyj (1940). Pivotally, Zheguc (2001, p. 8) provides the description of the phonetic features of the lects, which assists in unifying the transcription.

From the sociolinguistic point of view, Zheguc (2001) provides the best possible ground for tagging metadata. It contains standard sociolinguistic information about the speakers (age, occupation, level of education, name) and the dialectologist who had initially recorded the texts, Ivan Paňkevyč. This not only enables providing the research with more data for the variationist study, but possibly facilitates insight into how a particular scholar chooses to represent a non-standard variety (Saenko, 2018).

Thus, while the representation of the texts is not the most detailed among the other resources, the quality of the material is probably the best. The transcription transparency and the relative abundance of sociolinguistic information make the data useful for sociophonetic, sociolexical, and sociogrammatical studies alike and necessitate including this resource in the corpus.

4.3. Kuraszkiewicz (1963)

Kuraszkiewicz (1963) contains a short description of each of the East Slavic small territorial lect groups, as well as texts from the lects of these groups. Among these are several texts from the Transcarpathian lects, recorded and transcribed mostly at the beginning of the XX century (Kuraszkiewicz, 1963, pp. 124–128).

The possible digitisation of the texts from Kuraszkiewicz (1963) faces some significant issues. The transcription system here, on the contrary to Nakonečna and Rudnyc'kyj (1940), is not explained, and it is rather hard to make a proper correspondence between it and IPA, or even standard Ukrainian. There are also no signs of either a standard-based additional transcription or a translation into any other language. Given the lack of audio recordings, this is quite problematic. The crucial issue, however, is the lack of metadata, aside from the group (Lemko/Bojko/Hutsul) and the source of transcription, including the dialectologist's name. Kuraszkiewicz (1963) also generally provides the particular area where the recording took place. This facilitates some understanding on how different researchers represented different varieties; however, studying the varieties themselves is going to be significantly complicated.

While not exactly rich with metalinguistic and linguistic information, Kuraszkiewicz (1963) possesses one crucial advantage over all the other sources, namely, the geographical distribution of the lects. It covers at least two lects from each group, bringing much-needed diversity to the dataset.

5. Corpus Creation

5.1. Corpus Design

The main purpose of the corpus is to provide insights into the processes of variation and change that were going on in the Transcarpathian lects, as illustrated by the morphosyntactic properties of the texts. The most widely distributed format to represent the morphosyntactic properties is `.conllu`, so this is the main format of the corpus.

The transcriptions within the different sources are drastically different, so the research opts to unify them. The main transcription system is IPA. The main representation system for morphological tagging, however, is a standard Ukrainian phonetic/phonematic rendering, applied by both Nakonečna and Rudnyc'kyj (1940) and modern small territorial lects corpora of East Slavic languages, for instance, Goldin and Kryuchkova (2011) or Ryko and Spiricheva (2020). This system renders non-standard texts in something more resembling a neighbouring standard, but with faith-

ful depiction of key phonetic, morphological, lexical and syntactical features (von Waldenfels et al., 2014). This helps to present the corpus to the audience unfamiliar with the IPA conventions. Crucially, this system also facilitates more effective automatic processing. The corpus has two layers of tagging: morphosyntactic for each token (part-of-speech, morphological tags, lemma, syntactic features), and an additional linguistic one. When possible, the corpus also provides translations into other languages.

5.2. Manual Digitisation

Due to the specific system of transcription for the given lects, using OCR techniques of e-Scriptorium (Kiessling et al., 2019) or Transkribus (Kahle et al., 2017) proved to be impossible. Therefore, the performance of digitisation process is manual.

The workflow was the following:

- Manually type the material into a machine-readable form.
- In cases of Kuraszkiewicz (1963) and Zheguc (2001) it is necessary to add a standardised layer.
- Split the whole material into documents (especially relevant for Kuraszkiewicz (1963) and Zheguc (2001); Nakonečna and Rudnyc'kyj (1940) already contains splits by document.
- Split the document into texts (especially relevant for Nakonečna and Rudnyc'kyj (1940), which does not provide splits by texts within the documents).
- Split the texts into sentences (mostly done by the scholars).
- Perform manual word tokenisation.
- Put data in `.conllu` format.
- When required, provide additional information about the ASJP basic vocabulary list and named entities to the *misc* section of the token. In the *misc* section, there are also WF and TF fields, the latter providing the IPA-based transcription (if the original data contains phonetic transcription), and the former – normalised token (standard-based representation with removed diacritics).

5.2.1. Manual Digitisation Case Study: Kuraszkiewicz (1963)

The section outlines the workflow of transforming the original transcription from Kuraszkiewicz (1963) to IPA transcription and standardised Ukrainian orthography. Example 1 below shows the original rendering. It is accompanied by a glossed[7] version and a translation of the sentence.

(1) Hutsul (Kuraszkiewicz, 1963, p. 125)

Ta̯i̯		pip
ta̯i̯		pip-∅
and.so		priestling-NOM.SG
l̥edwy	u̯kik	
l̥edwy	u̯-kik-∅	
barely	away-run-PST.MASC.3SG	
witt'oho		gikka
wit-t-'oho		gikk-a
from-DET-MASC.GEN.SG		devil-GEN.SG

.
.
.

'And so the priestling barely ran away from that devil.'

The next step requires the conversion of this example into IPA and standardised Ukrainian, with the preservation of its main features. In this text, the most notable feature distinguishing Hutsul from Modern Standard Ukrainian is the sound change from the dental plosive *t̯* to the velar plosive *k* before palatal vowels or palatalised consonants. Two words exhibit this change: *u̯kik* 'ran away' (Modern Standard Ukrainian *ut^{j}'ik* 'id.') and *gikka* 'devil' (Modern Standard Ukrainian *'d^{j}ad^{j}ko* 'uncle'). As this appears to be an important feature of the lect, it should be preserved. At the same time, since the preposition від is generally a separate token in Modern Standard Ukrainian, the standardised version should separate it from the determiner т'ого. Thus, the preliminary reconstruction, based on other already processed texts from the corpus that have been standardised by the scholars who originally transcribed them, as well as on corpora of Modern Standard Ukrainian (Kopp et al., 2023), would be as follows: Тай пі́п ле́дви вкі́к від то́го гіккя.

IPA transcription should also rely on other texts and previous studies (Zilyns'kyj, 1932; Nakonečna and Rudnyc'kyj, 1940; Ševel'ov, 1979; Zhovtobriukh et al., 1979). Using this information, it is possible to reconstruct the following sound form of the sentence: taj p^{j}'ip ledvï wk^{j}'ik wit⌣t'oɦo gik^{j}k^{j}'a //. One of the key changes employed here is the placement of stress on the last syllable of gik^{j}k^{j}'a (by analogy with p'ip – pop'a). Another change is the substitution of *a* with a, which is characteristic of Hutsul in the GEN.SG of masculine nouns (Nakonečna and Rudnyc'kyj, 1940, p. 74). It is unclear what sound *l̥* designates; given other contexts in Kuraszkiewicz (1963) and the data from Nakonečna and Rudnyc'kyj (1940), the best option seems to be the IPA *l*. As the transcriber joined

[7] Glosses are given according to Comrie et al. (2008)

wit⌣tˈoño into a single sequence, the IPA transcription uses the ‿ symbol to indicate the pronunciation of two words without a pause. The IPA transcription also uses ʲ instead of an acute accent to indicate the palatalisation of consonants. Consequently, word stress is marked with the special symbol ˈ. The last crucial transformation is the use of *v* for original *w*, as the latter is most probably due to the transcriber using the Polish designation of this sound as part of the transcription, and *w* for original u̯, as *w* is a standard IPA designation of non-syllabic *u*. Other symbols coincide between the transcription in Kuraszkiewicz (1963) and the IPA. The only addition in this transcription is the / symbol marking the end of the sentence to denote a pause in the speaker's speech. After this conversion, the next step was to transform all the gathered data into the `.conllu` format. Table 2 shows the result.

```
# sent_id = chłopiec000.14
# IPA_transcription = tɑj pʲˈip ledvï wkʲˈik
wit⌣tˈoño gikʲkʲˈɑ //
# standard_text = Тай пíп лéдви вкíк від тóго
гíкькя.
# english_text = And so the priestling barely ran
away from that devil.

1 Тай _ _ _ _ _ _ _ wf="Тай"|tf="tɑj"

(...)
```

Table 2: The `.conllu` representation of `chłopiec000.14`. As the table is an illustration, it shows (for brevity) only the first token.

Repeating these steps for all sentences prepared the corpus for the next phases of tagging. These stages included automatic annotation, manual correction, and the provision of sociolinguistic information about the speakers.

5.3. Automatic Annotation

The code used for automatic annotation is available in an Open Science Framework (OSF) repository[8]. The annotation process follows the workflow described in the `README.md` file in the repository.

The main tool of automatic annotation is Stanza (Qi et al., 2020), a well-known set of pre-trained models designed to perform basic NLP tasks, providing output in the Universal Dependencies (UD) format (de Marneffe et al., 2021). Among others, the ones especially relevant for this study are: part-of-speech tagging, morphological tagging, lemmatisation, and dependency parsing. The automatic annotation pipeline includes running the Stanza model, trained on the standard Ukrainian corpus from UD (Nivre et al., 2020). This is due to the similarity of the selected representation system and the standard Ukrainian graphic system: it allows for the best possible results.

When having performed tagging, the script deletes XPOS, replacing them with underscore symbol: these are language-specific tags for standard Ukrainian. While, due to the high degree of similarity, some of them still can be applicable to the Transcarpathian lects, the differences within the grammatical systems of even rather closely related varieties are generally still significant enough to cause issues in the XPOS schema (Shishkina and Lyashevskaya, 2021). Thus, the decision was to temporarily remove this feature. After XPOS deletion, the data are ready for manual correction.

5.3.1. Note on the Use of GenAI

While Generative AI (GenAI) is extremely useful in low-resourced settings, when compared to the more traditional models (Baturova et al., 2025), its zero-shot application, even when there are high-resource closely related languages in the pre-training dataset, may still be problematic (Umbet et al., 2025). At this stage, there are not enough resources in the corpus to provide examples for GenAI prompts. For the further stages of the research, when the first texts for each Bojko, Hutsul and Lemko are fully digitised and cross-checked, the experiments with GenAI, as compared to the traditional tools, are necessary.

5.4. Manual Correction and Preliminary Analysis

Manual correction includes the editing of incorrect part-of-speech and syntactic feature tags, along with lemma assignment. The manual correction was carried out by a single annotator with previous expertise in annotating East Slavic languages.

A crucial stage of manual correction is error analysis. As the automatic tagging stage relies on Stanza applied cross-linguistically, there are many errors, especially those caused by errors in previous tagging phases. Table 3 shows the quantitative evaluation of each category using traditionally applied metrics; the following paragraphs provide a qualitative analysis.

5.4.1. Part-of-Speech/Morphological Tagging

The errata in part-of-speech tagging and, subsequently, morphological tagging emerge heavily from the differences in distributions of some items between standard Ukrainian and the Transcarpathian lects. In this fashion, Stanza tags То (t-o, DET-NEUT.NOM.SG) in LA1407.1.4 as a particle

[8]OSF link: https://osf.io/528zy/overview (last accessed: March 22, 2026).

Task	Metric	Score
Part-of-speech tagging	F1-score	66.78
Morphological tagging	Macro-F1 score	71.00
Lemmatization	Accuracy	75.7
Dependency parsing	UAS	67.65
Dependency parsing	LAS	52.71

Table 3: Evaluation of Stanza using traditionally applied metrics: F1-score for part-of-speech tagging, macro-F1 score for morphological tagging, accuracy for lemmatization, and unlabeled and labeled attachment scores (UAS and LAS) for dependency parsing.

(PART). While in standard Ukrainian то (to) is indeed a particle, то in the analysed text is rather a demonstrative pronoun (det), more akin to standard Ukrainian це (ʦʲ-e, DET-NEUT.NOM.SG).

This error also causes a subsequent chain of errors, the most prominent being the complete absence of the morphological tags required for the demonstrative pronoun. This underscores one of the crucial issues of Stanza-based pipelines: the simultaneous tagging of part-of-speech and morphology.

In other cases, the part-of-speech tag is correct, but some of the morphological tags are not. For instance, the model tags Фóльварк[footnotel] provide the form with stress for illustrative purposes. However, the forms that the model tagged underwent normalisation to exclude this factor of errata. (fˈolʲvark-∅'Folvark.village-NOM.SG') as NAME-TYPE=SUR (family name). This is especially frequent with some village names that get confused with family names or given names, which underscores the issues of out-of-domain tagging (Lyashevskaya and Afanasev, 2021).

5.4.2. Lemmatisation

The lemmatisation errata often stem from the incorrect part-of-speech tags. This is the case, for instance, of the model transforming вши́тки (fʂˈɯtkɣ, 'every-NOM.PL') to вшиток as a consequence of NOUN part-of-speech tag. In fact, вши́тки is an adjective, and therefore should get lemma вшиткий. The other cases include, for instance, lemmatising ме́нджі (mˌʤɻ'between' to менджь, triggered by NOUN tag.

The other type of error is the combination of lexical differences and the inability of the models to account for phonetic or morphological properties. Thus, Русна́ці (rusˈɑʦ-i'Rusnak-NOM.PL') becomes Руснац instead of expected Руснак. The phonetic changes that led to this ц/к alternation (Zhovtobriukh et al., 1979, pp. 119–120) are common for the Transcarpathian lects and standard Ukrainian, cf. роцi (rˈʦ-i, 'year-NOM.PL') – рiк. It is clear that the model does not grasp this kind of alternation.

In some cases, the combination of the lack of training material and the grammatical differences may also cause lemmatisation errata. For instance, ни́ма (n-ˈimɑ, 'they-INS.PL') is a form analogous to standard Ukrainian ними (n-imi, 'id.'; see Section 3.2). Its lemma is вони. The model, however, picks Кіма, which is a clear generation error, caused by the absence of both ними and the words ending with ма and being in INS.PL in the training dataset. This shows that Ukrainian is still a low-resourced language in terms of the UD corpora.

5.4.3. Dependency Parsing

The dependency parsing errata are multiple, but mostly have a single cause: incorrect part-of-speech tag. Thus, the aforementioned То gets DISCOURSE tag, while in fact it is NSUBJ. In case it were tagged as DET (as it should have been), there is a high chance that the assigned syntactic tag would have been correct.

5.4.4. Discussion and Final Representation

Overall, the errata made by the model stem not from significant differences in the morphological or syntactical structure of standard Ukrainian and the Transcarpathian lects, but rather from their lexical differences, low training material (Ukrainian corpus in UD 2.12, on which Stanza for Ukrainian was trained, has only 114 000 tokens), and running Stanza as a pipeline without manual checks in-between. Still, the manual check significantly reduced the errata of the model. Table 4 shows an example of a data piece after the manual check.

5.5. Metadata

The tagging of metadata, while not pivotal in terms of, for instance, automatic processing, is a critical part of both language variation studies in general and corpus-based dialectology in particular (Tagliamonte, 2025, pp. 109–111). The texts within the corpus are thus going to receive the metadata tag of belonging to one of these three groups, according to their descriptions within the sources, as well as the discovered phonetic features of the lects.

Where it is possible, the texts receive the tags of information about the speakers themselves. The first name and the last name undergo encryption via being transliterated into the Roman script, and afterwards abbreviated to the format "first letter of the first name + second name + the last three digits

sent_id = LA1407.1.4
IPA_transcription = to sut fʂ'ɯtkɣ lemk'ӧwskɣ s'elɑ de po lemk'ӧwskɣ fiw'ɑrjɑt
standard_text = То сут вши́тки лемки́вски се́ла, де по лемки́вски гва́рят
variation_text = {То=>Тото:::variation_type=Morph~Pronoun_Long} сут вши́тки лемки́вски се́ла, де по лемки́вски {г=>г:::variation_type=Phon~G}ва́рят.
german_text = Das sind alles lemkische Dörfer, wo lemkisch gesprochen wird.
english_text = These are all Lemko villages where Lemko is spoken.

1 То то PRON _ Animacy=Inan|Case=Nom|Gender=Neut|Number=Sing|PronType=Dem
2 nsubj 2:nsubj wf="То"|tf="то"

(...)

Table 4: The manually checked annotation of LA1407.1.4. As the table is an illustration, it shows (for brevity considerations) only the first token.

of the year of birth, otherwise 000". Thus, Василь Кабалюк[9], born in 1870, has id *vkabaluk870*, while Гафія Прумптула[10], information on whose year of birth is not in the data, has id *hprumptula000*. This is due to the variationist studies conventions and the consensus about the anonymity of the speakers[11] (Tagliamonte, 2025, pp. 48–50). Other features of the speakers include year of birth, place of birth (if known), education and occupation, where available.

More information is available from the recordings themselves. The corpus takes date and place of original publishing, genre and form (there are some songs in the dataset, so it is necessary to distinguish between poetry and prose). This enables the basic discourse analysis.

5.6. Representation and Access

The corpus is available on Zenodo as a separate language resource[12]. The results of the experiments are available in the OSF repository[13].

6. Current State: Conclusion

The article presents the design of a corpus of historical Transcarpathian lects. The study outlines the resources used for the creation of this corpus. It demonstrates the research pipeline utilised in building the corpus, from resource collection and manual digitisation to the manual correction of errors produced by the Stanza model. The main contribution is the open-access texts forming the Lemko part of the corpus. Table 5 demonstrates the main characteristics of the corpus.

Parameter	Value
Number of lects	3
Number of speakers	3
Number of texts	9
Number of sentences	90
Tagging status	Gold
Number of data collectors	2
Number of annotators	1
Token count	1949

Table 5: Key characteristics of the corpus; a more detailed overview is provided in the supplementary material.

The study provides an analysis of the errors produced by the Stanza model. It shows that the performance of the part-of-speech and morphological models is pivotal for all subsequent stages. The analysis also demonstrates that lexical differences are among the most crucial factors affecting model performance and that Stanza is not able—at least when trained on 110,000 tokens—to capture certain phonetic change patterns present in the dataset.

The next stage involves further expanding the corpus. Now that part of the material has been manually checked, the next step is to use GenAI either as an aid to the Stanza toolkit or as the main tagger. The corpus also requires English translations of the sentences for improved accessibility and, possibly, a TEI representation for a more qualitative study.

Limitations

The corpus by now is in its development phase, which means that not all the texts from the resources are present and tagged. One of the key

[9] Fictionalised name, not attested in the data.

[10] Fictionalised name, not attested in the data.

[11] While the original resources are available for research purposes and have a permissive license, they have not been widely accessed before the publication. Thus, it is better to adapt additional measures for privacy safety.

[12] Zenodo link: https://doi.org/10.5281/zenodo.19158682 (last accessed: March 22, 2026).

[13] OSF link: https://osf.io/528zy/overview (last accessed: March 22, 2026).

issues is that only a single annotator worked on the gold tagging, which could have introduced bias into it.

Ethical Considerations

The data has been published in the printed form and available for research purposes for fifty to ninety years by the time this article was written. Still, I anonymise the metadata, where possible, masking the names of the speakers, to compensate for possible ethics violations that could have happened at the time of the material collection.

The data themselves can contain slight mentions of xenophobic behaviour and religious (mostly, Christian) imagery. Discretion is advised.

Disclosure of Generative AI use

This study does not employ generative AI in the research process. While Stanza (Qi et al., 2020) technically belongs to the broader class of generative models in the modern colloquial meaning (the decoder models with more than a billion parameters trained on high-resource corpora), it operates at a much smaller scale and is locally reproducible. During the editing stage, the author used generative AI tools (Grammarly and OpenAI) solely for language polishing where non-native proficiency might otherwise have limited grammatical or stylistic clarity. The intellectual content of the article is entirely human-authored.

Acknowledgements

I thank the anonymous reviewers for their insightful feedback, which substantially improved the article. I am also grateful to the speakers whose recorded speech preserves the studied lects, to the scholars responsible for the original transcriptions, and to the research teams who produced the revised versions. Special thanks are due to Olha Fedorivna Mygolynets (ukr. Ольга Федорівна Миголинець, University of Uzhhorod) for her invaluable assistance with transcription systems and the phonetics of the analysed lects.

7. Bibliographical References

References

Ilia Afanasev. 2025. Computer-assisted study of historical Lemkian (Transcarpathian) lects: Basic vocabulary approach. *Scripta & e-Scripta*, 25:11–24.

Ilia Afanasev and Olga Lyashevskaya. 2023. From web to dialects: how to enhance non-standard Russian lects lemmatisation? In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 167–175, Gothenburg, Sweden. Association for Computational Linguistics.

Ludmila I. Barannikova. 2005. Govory territorij pozdnego zaselenija i problema ih klassifikacii [dialects of late-settled territories and the problem of their classification]. In Valentin E. Goldin and Olga Yu. Kryuchkova, editors, *Barannikova L. I. Obshhee jazykoznanie: izbrannye raboty [L. I. Barannikova. General linguistics: Selected works]*, pages 192–203. KomKniga.

Dari Baturova, Sarana Abidueva, Dmitrii Lichko, and Ivan Bondarenko. 2025. Low-resource buryat-Russian neural machine translation. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 85–93, Vienna, Austria. Association for Computational Linguistics.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. https://www.eva.mpg.de/lingua/resources/glossing-rules.php. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Salvatore Del Gaudio. 2017. *An introduction to Ukrainian dialectology*. Wiener slawistischer Almanach. Linguistische Reihe Sonderband 94. Peter Lang, Frankfurt am Main Bern Wien.

Valentin E. Goldin. 1990. K proektu tekstovogo dialektologicheskogo podfonda mashinnogo fonda russkogo jazyka [on the project of the textual dialectological sub-fund of the machine fund of the russian language]. In *Materialy III Vsesojuznoj konferencii po sozdaniju Mashinnogo fonda russkogo jazyka [Materials of the 3rd All-Union Conference on the Creation of the Machine Fund of the Russian Language]*, pages 92–103, Moscow. Izd-vo Moskovskogo universiteta.

Valentin E. Goldin and Olga Yu. Kryuchkova. 2011. Korpus russkoi dialektnoi rechi: kontseptsiia i parametry otsenki [Corpus of Russian Dialectal Speech: Concept and Evaluation Parameters]. In *Komp'uternaia lingvistika i intellektual'nye*

tekhnologii : Materialy ezhegodnoi Mezhdunarodnoi konferentsii, Bekasovo, 25–29 maia 2011 goda [Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference, Bekasovo, May 25–29, 2011], volume 10, pages 359–367, Moscow. Russian State University for the Humanities.

Irina B. Kachinskaya and Dmitrii V. Sichinava. 2015. Dialektnyj korpus segodnja [the dialect corpus today]. *Trudy Instituta russkogo jazyka im. V.V. Vinogradova*, 6:142–163.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Ljudmila Ė. Kalnyn'. 1973. *Opyt modelirovanija sistemy ukrainskogo dialektnogo jazyka: fonologicheskaja sistema [An attempt at modeling the system of the Ukrainian dialectal language: The phonological system]*. Nauka, Moscow.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. escriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.

Ezequiel Koile and George A. Moroz. 2024. Detecting linguistic variation with geographic sampling. *Journal of Linguistic Geography*, 12(1):24–31.

Vladimir B. Krys'ko. 1998. Drevnij novgorodsko-pskovskij dialekt na obshcheslavjanskom fone [the Old Novgorod-Pskov dialect against a common Slavic background]. *Voprosy Jazykoznanija*, 3:74–93.

Władysław Kuraszkiewicz. 1963. *Zarys dialektologii wschodniosłowiańskiej z wyborem tekstów gwarowych*, wyd. 2., zmien. i rozsz. edition. Państwowe Wydawn. Naukowe, Warszawa.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2025. Retrieval of parallelizable texts across Church Slavic variants. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 105–114, Abu Dhabi, UAE. Association for Computational Linguistics.

Olga Lyashevskaya and Ilia Afanasev. 2021. An hmm-based pos tagger for old church slavonic. *Journal of Linguistics/Jazykovedný casopis*, 72(2):556–567.

Ranko Matasović. 2008. *Poredbenopovijesna gramatika hrvatskoga jezika [The historical-comparative grammar of the Croatian language]*. Matica hrvatska, Zagreb. Biblioteka Theoria.

Aleksandra Miletić and Janine Siewert. 2023. Lemmatization experiments on two low-resourced languages: Low Saxon and Occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 163–173, Dubrovnik, Croatia. Association for Computational Linguistics.

George A. Moroz. 2016. Adverbial'nye konstrukcii vremennoj distribucii v balto-slavjanskih jazykah: areal'noe i korpusnoe issledovanie [adverbial constructions of temporal distribution in the balto-slavic languages: An areal and corpus study]. In N. Kazanskij and D. V. Gerasimov, editors, *Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskih issledovanij RAN (Tom XII, chast' 1) [Acta Linguistica Petropolitana. Proceedings of the Institute for Linguistic Studies of the Russian Academy of Sciences (Volume XII, Part 1)]*, volume XII/1, pages 151–167. Nauka, Saint Petersburg.

George A. Moroz. 2024. Skorost' russkoj rechi na osnove bilingval'nyh i dialektnyh ustnyh korpusov [the speed of russian speech based on bilingual and dialectal oral corpora]. In N. A. Korotaev and N. R. Sumbatova, editors, *Sostav nauki: Sbornik statej k jubileju Very Isaakovny Podleskoj [The Composition of Science: A Collection of Articles for the Anniversary of Vera Isaakovna Podleskaya]*, pages 366–378. Buki Vedi, Moscow.

Ol'ha F. Myholynec'. 2004. *Ukraïns'ki zakarpats'ki hovirky : teksty*. Lira, Užhorod.

Hanna Nakonečna and Jaroslav Bohdan Rudnyc'kyj. 1940. *Ukrainische Mundarten : Südkarpatoukrainisch ; (Lemkisch, Bojkisch und Huzulisch) [Ukrainian dialects: South Carpathian Ukrainian; Lemkian, Bojkian and Huzulian]*. Arbeiten aus dem Institut für Lautforschung an der Universität Berlin ; 9. Otto Harrassowitz, Berlin.

National Corpus of the Belarusian Language. 2018. Nacyjanalny korpus bielaruskaj movy ŭ kantekscie corpusnaj linhvistyki slavianskich krain [the national corpus of the belarusian language in the context of corpus linguistics of slavic countries]. In *XVI International Congress of Slavists*, Belgrade.

Tetjana V. Nazarova. 1977. *Hovory ukraïns'koï movy: zbirnyk tekstiv [Dialects of the Ukrainian language: A collection of texts]*. Naukova Dumka, Kyiv.

Sergei L. Nikolaev. 1988. Sledy osobennostei vostochnoslavianskikh plemennykh dialektov v sovremennykh velikorusskikh govorakh. 1. Krivichi [Traces of Features of East Slavic Tribal dialects in Modern Great Russian Dialects. 1. Krivichi]. In *Baltoslavianskie issledovaniia 1986 [Balto-Slavic Investigations 1986]*, pages 115–154. Nauka, Moscow.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4034–4043, Marseille, France.

Viktória Ondrejová and Marek Šuppa. 2024. Can LLMs handle low-resource dialects? a case study on translation and common sense reasoning in šariš. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 130–139, Mexico City, Mexico. Association for Computational Linguistics.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Achim Rabus and Yves Scherrer. 2017. Lexicon induction for spoken Rusyn – challenges and results. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 27–32, Valencia, Spain. Association for Computational Linguistics.

Anastasiia Ryko. 2024. A study of russian-belarusian border dialects: The use of the genitive case ending -u in the khislavichi dialect. *Zeitschrift für Slawistik*, 69(4):747–765.

Anastasiia I. Ryko and Margarita V. Spiricheva. 2022. The degree of preservation of dialectal features in different generations (khislavichi district of the smolensk region). *RSUH/RGGU Bulletin: "Literary Theory. Linguistics. Cultural Studies" Series*, 5:121–141. (In Russian).

Mikhail N. Saenko. 2018. Netochnosti v opisanii semantiki, vyzvannye vosprijatiem dialektnoj leksiki skvoz' prizmu literaturnogo jazyka: neskol'ko primerov [inaccuracies in the description of semantics caused by the perception of dialectal vocabulary through the prism of the literary language: Several examples]. In *Issledovanija po slavjanskoj dialektologii 19–20. Slavjanskie dialekty v sovremennoj jazykovoj situacii. Dialektnyj slovar' kak sposob issledovanija slavjanskix dialektov [Studies in Slavic dialectology 19–20. Slavic dialects in the modern language situation. Dialect dictionary as a method of studying Slavic dialects]*, pages 218–222. Institut slavjanovedenija RAN, Moscow.

Mikhail N. Saenko. 2020. Taxonomy of Slavic languages, history of the. In M. L. Greenberg, editor, *Encyclopedia of Slavic Languages and Linguistics Online*. Brill.

Yves Scherrer and Achim Rabus. 2017. Multi-source morphosyntactic tagging for spoken Rusyn. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 84–92, Valencia, Spain. Association for Computational Linguistics.

Yana Shishkina and Olga Lyashevskaya. 2021. Sculpting enhanced dependencies for belarusian. In *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, page 137–147, Berlin, Heidelberg. Springer-Verlag.

Sali A. Tagliamonte. 2025. *Analysing Sociolinguistic Variation*. Cambridge University Press.

Rudolf Trüb. 1989. Der Sprachatlas der deutschen Schweiz (SDS): ein Grossatlas für einen Kleinraum [the language atlas of german-speaking switzerland (sds): A large atlas for a small area]. In Werner H. Veith and Wolfgang Putschke, editors, *Sprachatlanten des Deutschen: laufende Projekte [Language Atlases of German: Ongoing Projects]*, pages 133–177. Niemeyer, Tübingen.

Sanzhar Umbet, Sanzhar Murzakhmetov, Beksultan Sagyndyk, Kirill Yakunin, Timur Akishev, and Pavel Zubitski. 2025. KazBench-KK: A cultural-knowledge benchmark for Kazakh. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 38–57, Vienna, Austria. Association for Computational Linguistics.

Ruprecht von Waldenfels, Michael Daniel, and Nina Dobrushina. 2014. Why standard orthography? building the ustya river basin corpus, an online corpus of a russian dialect. In *Komp'juternaja*

lingvistika i intellektual'nye technologii: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» (Bekasovo, 4 — 8 ijunja 2014 g.) [Computational Linguistics and Intellectual Technologies: Based on the materials of the Annual International Conference "Dialog" (Bekasovo, June 4-8, 2014)], volume 13, Moscow. Izd-vo RGGU.

Björn Wiemer, Ilja Seržant, and Aksana Erker. 2017. Convergence in the Baltic-Slavic contact zone: Triangulation approach. In Juliane Besters-Dilger, Cynthia Dermarkar, Stefan Pfänder, and Achim Rabus, editors, *Congruence in Contact-Induced Language Change*, pages 15–42. De Gruyter.

Björn Wiemer and Ilja A. Seržant. 2020. East Slavic dialectology: Achievements and perspectives of areal linguistics. In I. A. Seržant and B. Wiemer, editors, *Contemporary Approaches to Dialectology: The Area of North, Northwest Russian and Belarusian Vernaculars*, volume 13 of *Slavica Bergensia*, pages 11–80. John Grieg AS, Bergen.

Andrej Zaliznyak. 2004. *Drevnenovgorodskij dialekt [The Old Novgorodian Dialect]*. Jazyki slavjanskoj kul'tury, Moscow.

Svetlana S. Zemicheva. 2020. Ot abarma do jashhichishka: razrabotka leksikograficheskogo komponenta tomskogo dialektnogo korpusa [from *abarm* to *jashhichishek*: Development of the lexicographic component of the tomsk dialect corpus]. *Voprosy Leksikografii*, 18:98–117.

Ivan Zheguc. 2001. *Vybrani teksty z hucul's'koho hovoru v Zakarpatti [Selected texts from the Hutsul dialect in Transcarpathia]*. I. Zheguc, Munich.

Mikhailo Andriiovych Zhovtobriukh, Vitalii M. Rusanivs'kyi, and Vitaliy H. Skliarenko. 1979. *Istoriia ukraïns'koï movy. Fonetyka [History of the Ukrainian Language. Phonetics]*. Naukova dumka, Kyiv.

Ivan Zilyns'kyj. 1932. *Opis fonetyczny języka ukraińskiego*. Polska Akademia Umiejętności ‹Kraków› / Komisja Językowa: Prace 19. Nakładem Polskiej Akad. Umiejętności, Kraków.

Ivan M. Zilyns'kyj. 1933. *Karta ukraïns'kych hovoriv : z pojasnennjamy ; mirylo 1:4.000.000*. Praci Ukräins'koho Naukovoho Institutu 14. Ukraïns'kyj Naukovyj Instytut, Warszawa.

Jurij Volodymyrovyč Ševel'ov. 1979. *A historical phonology of the Ukrainian language*. Historical phonology of the Slavic languages ; 4. Winter, Heidelberg.

Language Resource References

Daniel, Michael and Dobrushina, Nina and von Waldenfels, Ruprecht. 2013–2018. *The Language of the Ustja River Basin: A Corpus of North Russian Dialectal Speech*. Linguistic Convergence Laboratory, NRU HSE.

Garder, M. O. and Petrova, N. S. and Moroz, A. B. and Panova, A. B. and Dobrushina, N. R. 2018. *Corpus of Spiridonova Buda Dialect*. Linguistic Convergence Laboratory, HSE University. Accessed on 24.10.2025.

Kopp, Matyáš and Kryvenko, Anna and Rii, Andriana. 2023. *Ukrainian parliamentary corpus ParlaMint-UA 4.0.1*. Slovenian language resource repository CLARIN.SI.

Marchenko, Igor A. and Dolgov, O. N. and Azanova, A. S. and Zambrzhitskaya, Maria S. and Zalivina, Ekaterina A. and Zemlyanskaya, S. A. and Mochul'skij, D. I. and Tsejtina, E. I. and Chistyakova, D. G. and Ron'ko, Roman V. 2025. *Database of the Dialectological Atlas of the Russian Language*. Institute of the Russian Language RAS. Accessed October 24, 2025.

Ronko, Roman and Volf, Elena and Grebyonkina, Maria and Ershova, Maria and Okhapkina, Anna and Khadasevich, Anna and Morozova, Valeria. 2019. *Corpus of Opochetsky Dialects*. Linguistic Convergence Laboratory, HSE University; V.V. Vinogradov Russian Language Institute Russian Academy of Science. Accessed on 24.10.2025.

Ryko, Anastasiia I. and Spiricheva, Margarita V. 2020. *Corpus of the Russian Dialect Spoken in Khislavichi District*. Linguistic Convergence Laboratory, HSE University. Available online at `https://lingconlab.ru/khislavichi/`, accessed on 23.10.2025.

Shvedova, Maria and von Waldenfels, Ruprecht and Yaryhin, Sergey and Rysin, Andriy and Starko, Vasyl and Nikolaenko, Tymofij and Lukashevskyi, Arsenii and others. 2017–2025. *General'nyj rehional'no anotovanyj korpus ukraïns'koï movy (HRAK) [General Regionally Annotated Corpus of the Ukrainian Language (GRAC)]*. University of Jena.

Ter-Avanesova, A. V. and Balabin, F. A. and Dyachenko, S. V. and Malysheva, A. V. and Panova, A. B. and Morozova, V. A. 2019. *Corpus of the Malinino Dialect*. Linguistic Convergence Laboratory, NRU HSE; V.V. Vinogradov Russian Language Institute of the Russian Academy of Science. Accessed on 24.10.2025.

Ter-Avanesova, A. V. and Dyachenko, S. V. and Kolesnikova, E. V. and Malysheva, A. V. and Ignatenko, D. I. and Panova, A. B. and Dobrushina, N. R. 2018. *Corpus of Rogovatka Dialect*. Linguistic Convergence Laboratory, NRU HSE. Accessed on 24.10.2025.

Wenker, Georg and Maurmann, Emil and Wrede, Ferdinand. 1889–1923. *Sprachatlas des Deutschen Reichs [Language Atlas of the German Empire]*. Research Center Deutscher Sprachatlas. Original manuscript (1889–1923). Published as the Digital Wenker Atlas (DiWA).

Zemicheva, S. S. and Dubtsova, L. A. and Gromov, M. L. and Galanina, V. V. and Ugryumova, M. M. and Vasilchenko, A. A. and Parshina, A. V. and Popova, D. P. and Duminskaya, A. V. and Zyuzkova, N. A. and Bukhanova, E. D. 2023. *Tomskij dialektnyj korpus 2.0 [Tomsk Dialect Corpus 2.0]*. Laboratorija obshhej i sibirskoj leksikografii NI TGU [Laboratory of General and Siberian Lexicography, National Research Tomsk State University]. Access mode: free.

The Generator-Eraser Paradox: Community Guidelines for Responsible LLM-Assisted Dialect Resource Creation

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Dialect resources occupy a unique position at the intersection of scientific description, cultural preservation, and computational infrastructure. Large language models offer powerful capabilities for accelerating dialect resource development through retrieval-grounded drafting, corpus navigation, metadata enrichment, and annotation workflow support. However, the same systems pose substantial risks: they can contribute to dialect erasure by privileging prestige varieties, homogenizing orthography, and enabling synthetic feedback loops that reduce linguistic diversity over time. These risks are particularly acute for language varieties characterized by diglossia, limited written standardization, or marginalized speaker communities. This paper makes three contributions. First, we integrate insights from variationist sociolinguistics and corpus linguistics to formalize the **generator-eraser paradox** as a theoretical framework for understanding the dual nature of LLM-assisted dialect work. Second, we derive **12 community guidelines** that operationalize this framework into implementable design requirements for dialect resource creation and documentation. Third, we provide an **in-depth case study of Arabic dialects**, including a structured comparison of widely used resources, to demonstrate how these guidelines address language-specific challenges including diglossia, orthographic variability, and community governance. The contribution is conceptual and operational rather than experimental, with the goal of enabling dialect communities and resource builders across languages to adopt LLMs without sacrificing authenticity, variation, or sovereignty.

Keywords: dialect resources, language variation, LLM governance, orthographic normalization, participatory design, community guidelines, Arabic dialects

1. Introduction

Language variation is not noise; it is the systematic mapping of linguistic forms to social meanings [Labov, 1972, Eckert, 2012]. Dialects, regional varieties, and sociolects encode community identity, geographic affiliation, social positioning, and cultural knowledge in ways that standard varieties cannot capture. For computational linguistics, dialect resources serve multiple functions: they enable training of dialect-aware NLP systems, provide evaluation benchmarks for measuring system fairness, support language documentation and preservation efforts, and create infrastructure for communities to engage with technology on their own terms.

Yet dialect resources remain chronically underdeveloped relative to standardized varieties. The resource gap documented by Joshi et al. [2020] applies not only across languages but within languages, where prestige varieties dominate corpora, evaluation sets, and model training data. This imbalance has measurable consequences. Automatic speech recognition systems exhibit higher error rates for speakers of African American English [Koenecke et al., 2020]. Hate speech detection systems disproportionately flag dialect features as toxic [Sap et al., 2019]. Part-of-speech taggers trained on standard varieties degrade when applied to dialect text [Hovy and Søgaard, 2015, Blodgett et al., 2016]. The common thread is that systems optimized for prestige varieties treat dialect features as deviations rather than as systematic linguistic patterns.

Large language models introduce both opportunity and risk into this landscape. On one hand, LLMs can accelerate resource creation: they can assist in drafting lexicographic entries from retrieved attestations, propose candidate annotations for human confirmation, generate pedagogical examples grounded in dialect corpora, and help navigate large collections of unstructured dialect data. On the other hand, LLMs are trained on corpora that reflect and amplify existing resource imbalances. Their statistical priors favor high-frequency forms, which in language contexts typically means prestige varieties. When deployed in resource creation workflows, they risk producing outputs that drift toward dominant norms, homogenize orthographic variation, and introduce synthetic material that, if recursively incorporated into future training, can narrow linguistic distributions and eliminate precisely the tail phenomena that constitute dialect distinctiveness [Shumailov et al., 2024].

This paper addresses a gap in the literature: the absence of a systematic framework for governing LLM involvement in dialect resource work. Our contributions are threefold:

1. We develop the **generator-eraser paradox** as a theoretical framework that synthesizes variationist sociolinguistics, corpus linguistics, and recent findings on model collapse to articulate

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 209–220
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

the dual nature of LLM assistance in dialect contexts.

2. We derive **12 community guidelines** from this framework, written as implementable design requirements with explicit checks and language-specific considerations.
3. We provide an **in-depth case study of Arabic dialects**, demonstrating how the guidelines address challenges of diglossia, orthographic variability, and multi-stakeholder governance that generalize to other diglossic and pluricentric language situations.

The paper is deliberately non-experimental. We do not report model performance metrics, and we avoid unverifiable claims about system behavior. Instead, we translate stable findings from sociolinguistics, corpus linguistics, and AI governance into operational requirements for responsible dialect resource development.

2. Background

2.1. Dialect Variation as Structured Diversity

Three waves of variationist sociolinguistics have established that dialect variation is systematic rather than random [Eckert, 2012]. The first wave, pioneered by Labov [1972], documented correlations between linguistic variables and social categories such as class, age, and gender. The second wave examined local communities and social networks, revealing how dialect features index local identity and group membership. The third wave reframed variation as a resource for constructing social meaning, emphasizing speaker agency in deploying dialect features for communicative purposes [Chambers, 2013].

For resource creation, these findings imply that dialect corpora are not merely collections of sentences; they are archives of variation. The resource object includes linguistic forms and their frequencies, the contexts and registers in which forms appear, metadata about speakers and communities where ethically available, and the orthographic practices and writing conventions that speakers employ. Any processing pipeline that alters these dimensions without traceability risks severing the link between form and social meaning.

2.2. Diglossia and Prestige Hierarchies

Many language situations involve diglossia: a stable relationship between a high variety associated with formal contexts and low varieties associated with everyday speech [Ferguson, 1959]. Arabic exemplifies classical diglossia, with Modern Standard Arabic occupying institutional domains while regional dialects dominate spoken interaction [Holes, 2004, Versteegh, 2014]. However, diglossia is not unique to Arabic. Similar configurations appear in Swiss German, Haitian Creole, Tamil, and many other language situations where written standards diverge from spoken vernaculars.

Diglossia creates a prestige gradient that shapes both human behavior and computational systems. When models or preprocessing pipelines attempt to normalize text, they tend to normalize toward the institutionally privileged variety. In resource work, such normalization can be an explicit design choice, such as adopting a conventional orthography, or an implicit side effect of preprocessing steps such as spell-checking, tokenization, or automatic correction [Eisenstein, 2013].

2.3. Corpus Representativeness and Provenance

Corpus linguistics distinguishes size from representativeness. A corpus must be designed or sampled to reflect intended domains, registers, and speaker groups, and it must include metadata enabling audit and interpretation [Biber, 1993, McEnery and Hardie, 2012]. Dialect corpora are frequently convenience samples from social media platforms with strong demographic and topical skews. These skews may favor urban over rural speakers, younger over older generations, particular countries or regions with higher platform penetration, and specific genres such as political commentary or entertainment.

Provenance matters because dialect variation is socially indexed. Without provenance, a resource cannot be reliably interpreted as a proxy for a dialect community; it becomes a proxy for platform use. Documentation frameworks such as datasheets for datasets [Gebru et al., 2021] and data statements [Bender and Friedman, 2018] provide templates for recording this information, but adoption remains uneven in dialect resource work.

2.4. LLM Risks: Representational Harm and Synthetic Recursion

Two distinct bodies of research document LLM risks relevant to dialect resources.

First, work on representational harms shows that language models reproduce and amplify biases present in training data [Blodgett et al., 2020, Bender et al., 2021]. In dialect settings, this includes bias toward prestige norms, stereotyped associations between dialect features and social categories, and performance degradation on dialect text that cascades through downstream applications [Jurgens et al., 2017, Tan et al., 2020].

Second, research on model collapse demonstrates that training on recursively generated data can eliminate low-probability events, leading to progressively impoverished distributions [Shumailov et al., 2024]. For dialect resources, low-probability events often correspond to precisely what makes dialects distinctive: minority lexical items, region-specific constructions, and socially marked forms that occur infrequently in any given corpus but carry substantial social meaning.

2.5. Existing Dialect Resources and Shared Tasks

The NLP community has developed substantial infrastructure for dialect processing. The VarDial workshop series and associated shared tasks have produced benchmarks for dialect identification across multiple language families [Zampieri et al., 2017]. For Arabic specifically, resources include the Arabic Online Commentary Dataset [Zaidan and Callison-Burch, 2011], the MADAR multi-dialect corpus and lexicon [Bouamor et al., 2018], the Gumar Gulf Arabic corpus [Khalifa et al., 2016], and shared tasks such as NADI [Abdul-Mageed et al., 2021]. Multi-dialect resources have also been developed for tasks such as author profiling [Zaghouani and Charfi, 2018], fine-grained dialect annotation [Charfi et al., 2019], and hate speech detection across Arabic varieties [Charfi et al., 2024a, Zaghouani et al., 2024]. For African American English, TwitterAAE provides a large-scale resource with demographic annotations [Blodgett et al., 2016]. Earlier surveys of freely available Arabic corpora [Zaghouani, 2014] have documented the landscape of accessible resources, revealing both the breadth and the gaps in existing coverage.

These resources demonstrate mature methodological practices but also reveal fragmentation. Different datasets adopt different orthographies, annotation frameworks, and access conditions. This fragmentation creates barriers to scalable dialect resource development and cross-resource interoperability.

3. The Generator-Eraser Paradox

We propose the **generator-eraser paradox** as a framework for understanding the dual nature of LLM involvement in dialect resource work. The paradox captures a fundamental tension: the same properties that make LLMs useful for resource creation also make them dangerous for dialect preservation.

3.1. The Dual Nature of LLM Assistance

LLMs can accelerate dialect resource development through several mechanisms. *Retrieval-grounded drafting* allows LLMs to draft lexicographic entries, example sentences, or educational materials grounded in retrieved dialect attestations; retrieval-augmented generation [Lewis et al., 2020] reduces hallucination and anchors outputs in authentic dialect data. *Corpus navigation and clustering* helps resource builders explore large unstructured collections, identify patterns, and cluster similar items for batch annotation. *Metadata enrichment* enables LLMs to propose metadata such as topic labels, register classifications, or dialect sub-variety tags for human confirmation. *Annotation workflow support* allows LLMs to generate candidate annotations that human experts review, revise, and approve, reducing annotation time while maintaining human authority over final decisions.

However, the same LLMs can erode dialect distinctiveness through countervailing mechanisms. *Prior-driven normalization* occurs because LLMs trained predominantly on prestige varieties have statistical priors that favor standard forms; without explicit constraints, generated outputs drift toward these priors. *Orthographic homogenization* reduces the variability that reflects phonological differences, community conventions, or the absence of written standardization. *Synthetic recursion* occurs when LLM-generated content enters training pipelines for future models or becomes part of dialect corpora without adequate labeling; model collapse research shows this effect is particularly severe for low-frequency events [Shumailov et al., 2024]. *Prestige drift*, which we define as the systematic substitution of dialect-specific markers (phonological, morphological, lexical, or syntactic) with high-variety equivalents over iterative processing, may be imperceptible in any single interaction but cumulative over project lifecycles.

3.2. Amplifying Factors

Several factors amplify the paradox in particular language situations. In *diglossic* settings, the high variety dominates written corpora and institutional language, creating severe training data imbalances that LLMs inherit and reproduce unless explicitly constrained. Dialects with *limited standardization* exhibit higher orthographic variability and less consistent lexicographic documentation; LLMs may treat this variability as error rather than as authentic variation. *Platform sampling* from social media introduces demographic, topical, and temporal skews that LLMs trained on such data may amplify. *Community power differentials* mean that dialect communities often lack institutional power to shape how their language is documented and processed; without participatory governance, external resource builders may make decisions that do not reflect community priorities.

4. Design Requirements

This section converts the theoretical framework into explicit design requirements that the proposed guidelines must address.

Preserving Distributional Properties. Dialect resources must preserve the distributional properties of authentic language use, including the frequency and context of variant forms, the presence of low-frequency items, and the orthographic diversity of written representations.

Layered Orthographic Representation. Orthographic normalization should be implemented as a reversible transformation with documented mappings, not as a destructive rewrite.

Transparent Provenance. All resource artifacts should include clear provenance information distinguishing human-produced content from LLM-assisted content.

Human Authority. Human experts, particularly native speakers and community members, must retain authority over final resource decisions.

Community Governance. Dialect communities should participate in scope setting, annotation framework design, release decisions, and harm review [Costanza-Chock, 2020, Bird, 2020].

Recursion Avoidance. Synthetic data must be transparently labeled and must not enter training pipelines or corpus additions without explicit evaluation of diversity impacts.

5. The 12 Community Guidelines

Each guideline is written as an implementable requirement with rationale, checks, and considerations for language-specific adaptation.

5.1. Guideline 1: Real Data Primacy

Requirement. Human-produced dialect data remains the primary substrate for all resource artifacts. Synthetic text can only serve as an explicitly labeled augmentation layer, never as a replacement for authentic data.

Rationale. Model behavior follows training distributions and resource composition [Bender et al., 2021]. Synthetic recursion can eliminate tail events that constitute dialect distinctiveness [Shumailov et al., 2024].

Checks. Maintain a datasheet documenting collection methods, coverage, and limitations [Gebru et al., 2021]. Store synthetic content separately with full provenance metadata including prompts, models, and generation parameters.

Adaptation. For languages with minimal written dialect tradition, spoken transcripts may constitute primary data, with transcription conventions documented.

5.2. Guideline 2: Retrieval Before Generation

Requirement. For resource drafting tasks, prefer retrieval-augmented generation over unconstrained prompting. Outputs should cite or derive from authenticated dialect attestations.

Rationale. Retrieval-augmented generation conditions outputs on evidence rather than priors [Lewis et al., 2020], reducing drift toward prestige norms.

Checks. Log retrieval sources for generated content. Implement verification that generated forms appear in the retrieval set. Monitor retrieval quality through: (1) coverage metrics measuring what proportion of dialect marker types in outputs are attested in retrieved sources; (2) recall for low-frequency markers that are most vulnerable to prestige substitution; and (3) source diversity ensuring outputs draw from multiple attestations rather than over-relying on single sources.

Adaptation. For under-documented dialects, retrieval sets may be small; in such cases, additional human validation compensates for limited grounding. When retrieval fails to return sufficient dialect-specific attestations, flag outputs for mandatory expert review rather than allowing fallback to unconstrained generation.

5.3. Guideline 3: Orthographic Preservation

Requirement. Preserve raw orthographic forms as the archival layer. If normalization is required for processing, implement it as a reversible transformation with documented mappings.

Rationale. Orthography encodes phonological, social, and identity information. Conventional orthography frameworks such as CODA for Arabic [Habash et al., 2012] provide principled approaches to reversible normalization.

Checks. Maintain three layers: raw orthography, normalized orthography, and the mapping between them. Document what information is lost or ambiguous in normalization. Mappings should specify: (1) the normalization rule applied; (2) whether the mapping is one-to-one (fully reversible) or many-to-one (lossy); and (3) the linguistic rationale. For example, a Gulf Arabic raw form *chinna* normalized to CODA *kinna* would be documented as: `raw: chinna → norm: kinna | rule: ch→k | type: lossy (ch/k merger) | note: ch reflects /tš/ pronunciation.`

Adaptation. For languages with Latin script, normalization may address capitalization and diacritics; for non-Latin scripts, language-specific conventions apply.

5.4. Guideline 4: Dialect-Aware Prompting

Requirement. Prompts must explicitly specify dialect identity and instruct against implicit translation to prestige varieties. Prompts should request uncertainty marking rather than confident guessing.

Rationale. LLMs default to dominant norms under ambiguity. Explicit constraints can reduce prestige drift.

Checks. Include dialect specification in all prompts used for resource creation. Log prompts with outputs for audit.

Adaptation. Dialect names and boundaries vary by language; use community-preferred terminology.

5.5. Guideline 5: Human Validation Gate

Requirement. No LLM-assisted resource artifact is released without review by native speakers and domain experts. Review covers orthographic authenticity, semantic accuracy, pragmatic appropriateness, and cultural acceptability.

Rationale. Representational harms are systematic and unevenly distributed [Blodgett et al., 2020]. Only community-grounded review can detect subtle errors that automated metrics miss.

Checks. Define review rubrics. Record reviewer judgments and disagreements as signals of variation rather than errors to resolve.

Adaptation. Reviewer selection should prioritize diversity across dialect sub-varieties, speaker demographics, and expertise types.

5.6. Guideline 6: Synthetic Data Governance

Requirement. Synthetic data must be transparently labeled in all resource artifacts and must not be incorporated into recursive training pipelines without controlled evaluation of diversity impacts.

Rationale. Recursive training on synthetic data causes model collapse and diversity loss [Shumailov et al., 2024].

Checks. Implement content-level provenance using structured metadata fields (e.g., in JSON/JSONL schemas: `is_synthetic: boolean`, `generation_model: string`, `generation_date: timestamp`, `prompt_hash: string`). Maintain separate partitions for human and synthetic content. For large-scale pipelines, consider dataset-level synthetic ratio caps (e.g., $\leq$20% synthetic content) with periodic contamination audits.

Adaptation. For shared tasks, clarify whether synthetic augmentation is permitted and require participants to report synthetic content ratios and generation methods.

5.7. Guideline 7: Diversity Monitoring

Requirement. Track whether resource releases retain dialect markers and do not exhibit drift toward prestige varieties or dominant regional norms.

Rationale. Representativeness and balance require monitoring, not assumption [Biber, 1993, McEnery and Hardie, 2012].

Checks. Recommended approaches include: (1) compiling dialect marker inventories (diagnostic lexemes, morphological patterns, phonological spellings) validated by community experts; (2) computing marker frequency distributions across resource versions; (3) measuring distributional divergence using Jensen-Shannon divergence or similar metrics between current and baseline versions; (4) tracking type-token ratios for dialect-specific vocabulary; and (5) applying mixed-effects models controlling for topic and register when comparing releases. Flag statistically significant drift for manual review.

Adaptation. Marker selection should reflect language-specific features that distinguish dialects; community input is essential for validating marker inventories and setting acceptable drift thresholds.

5.8. Guideline 8: Participatory Governance

Requirement. Dialect communities participate in defining resource scope, annotation frameworks, release conditions, and harm review processes.

Rationale. Participatory design supports community self-determination and reduces extractive dynamics [Bird, 2020, Costanza-Chock, 2020].

Checks. Document community consultation processes. Include community representatives in project governance.

Adaptation. Community structure varies; some dialects have organized advocacy groups, others require identifying representative speakers through sociolinguistic fieldwork.

5.9. Guideline 9: Data Sovereignty and Access Control

Requirement. Clarify ownership and access rights for all resource components. Adopt tiered access when community safety, privacy, or cultural sensitivity requires restriction.

Rationale. Some dialect communities face political risk, discrimination, or cultural appropriation concerns. This guideline aligns with the CARE Principles for Indigenous Data Governance [Carroll et al., 2020].

Checks. Specify licensing terms. Document rationale for access restrictions. Consider benefit-sharing arrangements that return value to contributing communities.

Adaptation. Legal frameworks vary by jurisdiction; consult local expertise on data protection and indigenous language rights where applicable [Rainie et al., 2019].

5.10. Guideline 10: Pipeline Documentation

Requirement. Publish a pipeline card documenting all preprocessing steps, orthography handling, prompting strategies, retrieval configurations, model versions, and validation procedures.

Rationale. Transparency enables audit, reproduction, and identification of potential harm sources [Gebru et al., 2021, Mitchell et al., 2019].

Checks. Use structured documentation templates. Version control pipeline configurations.

Adaptation. Documentation should be accessible to both technical and community audiences; plain language summaries complement technical specifications.

5.11. Guideline 11: Prestige Drift Audits

Requirement. Conduct periodic audits to detect systematic replacement of dialect-specific forms with prestige variety equivalents.

Rationale. Drift may be imperceptible in individual outputs but cumulative over project lifecycles.

Checks. Audit procedures should include: (1) random sampling of outputs for manual expert review against dialect authenticity criteria; (2) automated comparison of dialect marker frequencies against baseline corpora using the metrics from Guideline 7; (3) targeted review of high-risk categories (e.g., function words, verbal morphology, discourse markers) where prestige substitution is most common; and (4) documentation of detected drift with root cause analysis (prompt design, retrieval gaps, or model limitations). Establish correction workflows for identified issues before release.

Adaptation. Audit frequency should reflect project pace; high-throughput workflows require more frequent checks, potentially automated continuous monitoring with periodic manual validation.

5.12. Guideline 12: Long-Term Stewardship

Requirement. Treat dialect resources as living infrastructure with periodic review, correction mechanisms, and versioned releases.

Rationale. Languages evolve, communities' needs change, and resource harms may emerge over time.

Checks. Establish maintenance schedules. Provide error reporting channels. Document version histories with changelogs.

Adaptation. Sustainability planning should consider institutional support, community capacity, and succession planning for key personnel.

6. Case Study: Arabic Dialects

Arabic provides a compelling case study because it exhibits the full range of challenges the guidelines address: classical diglossia, substantial orthographic variability, multiple regional varieties with distinct features, and complex political and cultural dynamics. This section demonstrates how the guidelines apply to Arabic while extracting generalizable lessons.

6.1. Diglossia, Orthography, and the Arabic Script System

Arabic diglossia involves Modern Standard Arabic (MSA) as the high variety used in formal writing, education, and media, alongside regional dialects that serve as primary spoken languages across the Arab world [Ferguson, 1959, Holes, 2004]. Subsequent scholarship emphasizes a continuum of registers and mixed styles, including extensive code-switching between MSA and dialect elements [Badawi, 1973]. For resource creation, this continuum means that dialect corpora frequently contain MSA elements, and annotation frameworks must decide how to represent mixture without forcing artificial binaries. The prestige gradient associated with diglossia shapes both human linguistic behavior and computational system outputs; when models or pipelines normalize, they tend to normalize toward MSA.

Arabic dialect writing occurs in Arabic script, Latin script (Arabizi), and mixed forms. Even within Arabic script, writers differ substantially in their orthographic choices. Several features of Arabic script are particularly relevant for understanding why orthographic variability is so pervasive in dialect writing and why preservation matters for resource creation (for a comprehensive treatment, see Habash 2010):

- **Hamza placement**: The glottal stop (*hamza*) can be written on different carriers (alif, waw, ya) or standalone, with placement rules that differ between MSA and dialect conventions. Dialect writers often omit or misplace hamza according to local pronunciation, creating systematic variation that carries phonological information.
- **Ta marbuta vs. ha**: The feminine marker *ta marbuta* () is sometimes written as *ha* () in dialectal text, reflecting pronunciation patterns. This

alternation is phonologically motivated in many dialects and should not be treated as a spelling error.

- **Alif variants**: The initial alif can appear with or without hamza, and with hamza above or below (vs. vs.). Dialect writers frequently use bare alif regardless of MSA rules, reflecting actual pronunciation.
- **Dialect-specific phonology in script**: Certain sounds that exist in dialects but not in MSA require creative orthographic solutions. For example, Gulf Arabic speakers may write the affricated /tš/ as *ch* in Arabizi or use the letter *kaf* () in Arabic script, creating many-to-one mappings that complicate normalization.

The CODA framework proposes a conventional orthography designed to reduce sparsity while preserving dialect identity [Habash et al., 2012], and CODA* extends these principles across multiple dialects [Habash et al., 2018]. CODA provides exactly the kind of reversible normalization layer that Guideline 3 requires; resources should maintain raw orthography as an archival layer, with CODA normalization applied as a separate, documented transformation.

6.2. Resource Landscape

Table 1 compares major Arabic dialect resources across dimensions relevant to LLM-assisted workflow design. This comparison highlights both the strengths of existing resources and the gaps that new projects must address.

6.3. Applying the Guidelines: Practical Strategies

Several guidelines have particular relevance for Arabic, and we outline both the challenges and concrete strategies for addressing them.

Guideline 1 (Real Data Primacy): Arabic dialect corpora should explicitly distinguish human-produced content from any synthetic augmentation. Given MSA dominance in training data, synthetic Arabic text is particularly prone to MSA drift. **Practical strategy:** Implement metadata fields at the sentence level marking provenance (human vs. synthetic), and enforce separation in storage. For shared task organizers, require participants to declare synthetic data usage and report synthetic content ratios.

Guideline 2 (Retrieval Before Generation): Retrieval from authenticated dialect sources reduces drift toward MSA synonyms. **Practical strategy:** Build retrieval indices over authenticated dialect corpora such as MADAR's parallel lexicon for lexicographic work, or Gumar for Gulf-specific vocabulary. Configure retrieval to prioritize dialect-specific sources and flag outputs where retrieval coverage falls below a defined threshold (e.g., fewer than 60% of generated dialect markers attested in retrieved sources).

Guideline 3 (Orthographic Preservation): CODA provides a principled normalization framework, but raw orthography must be maintained as the archival layer. **Practical strategy:** Store resources in a three-layer format: raw user text, CODA-normalized text, and a structured mapping log that records each normalization decision along with its type (reversible or lossy) and linguistic rationale. For the Arabic-specific features described above (hamza, ta marbuta, alif variants), the mapping log should flag instances where normalization collapses phonologically meaningful distinctions. Existing tools such as CAMeL Tools [Obeid et al., 2020] can support automated components of this pipeline.

Guideline 4 (Dialect-Aware Prompting): Prompts must specify the target dialect using community-preferred terminology (e.g., Egyptian, Tunisian, Gulf) and explicitly instruct against MSA substitution. **Practical strategy:** Develop prompt templates that include (a) explicit dialect specification, (b) instructions to preserve dialect-specific lexical items and morphological patterns, (c) requests to flag uncertainty rather than defaulting to MSA forms, and (d) example dialect-specific outputs to anchor generation. For instance, when generating Tunisian Arabic content, prompts should specify that negation uses *ma...sh* rather than MSA *la/lam*, and that future tense uses *besh* rather than MSA *sawfa*.

Guideline 7 (Diversity Monitoring): Arabic's rich dialectal variation requires monitoring along multiple dimensions simultaneously. **Practical strategy:** For each target dialect, compile a validated inventory of diagnostic markers spanning phonological spellings (e.g., Gulf *ch* for /tš/), morphological patterns (e.g., Egyptian *bi-* prefix for present tense), lexical items (e.g., Maghrebi *barsha* for "a lot"), and syntactic constructions. Track the frequency of these markers across resource versions and flag declines exceeding community-defined thresholds for manual review.

Guideline 8 (Participatory Governance): Arabic-speaking communities span multiple countries with distinct political contexts. **Practical strategy:** Establish advisory boards that include speakers from diverse dialect regions, balancing representation across Mashreqi and Maghrebi varieties, urban and rural speakers, and different age groups. For pan-Arabic resources, governance should involve representatives from multiple countries to avoid privileging any single national variety.

Guideline 9 (Data Sovereignty): Some dialect communities face political sensitivities related to

Resource	Coverage	Domain	Guideline Implications	LLM Workflow Role
AOC [Zaidan and Callison-Burch, 2011]	Multiple dialects	News comments	Retrieval grounding (G2); privacy and access control (G9)	Retrieval source for lexicographic work; dialect identification training
MADAR [Bouamor et al., 2018]	25 city dialects	Travel domain	RAG grounding (G2); representativeness monitoring (G7)	Parallel lexicon for cross-dialect retrieval; generation anchoring
Gumar [Khalifa et al., 2016]	Gulf Arabic	Forum novels	Gulf-specific retrieval (G2); drift detection (G11)	Gulf variety benchmarking; orthographic variation archive
QADI [QCRI, 2020]	18 countries	Tweets	Reproducibility (G10); documentation (G10)	Country-level dialect classifier training
NADI [Abdul-Mageed et al., 2021]	Country/province	Tweets	Nuanced labels (G7); privacy (G9)	Fine-grained dialect identification; evaluation benchmark
CODA* [Habash et al., 2012, 2018]	Cross-dialect orthography	Standardization	Reversible normalization (G3); layered representation	Normalization pipeline component; orthographic mapping reference
BOLT [Linguistic Data Consortium, 2021]	Egyptian Arabic	SMS/Chat	Evaluation (G5); tiered access (G9)	Egyptian variety benchmark; informal register source
AraP-Tweet [Zaghouani and Charfi, 2018]	16 countries, 11 regions	Tweets	Demographic balance (G7); diversity (G8)	Multi-dialect profiling; dialect-demographic correlation
ADHAR [Charfi et al., 2024a]	MSA + 4 dialect groups	Tweets	Multi-dialect balance (G7); harm review (G8)	Dialect-aware hate speech detection; bias auditing

Table 1: Arabic dialect resources with guideline implications and LLM workflow roles.

regional identity or minority status. **Practical strategy:** Implement tiered access for resources containing location-identifiable content, with stricter controls for data from politically sensitive regions. Consider removing or anonymizing fine-grained geographic metadata when full access is granted.

Guideline 11 (Prestige Drift Audits): Audits should monitor for MSA substitution and for drift toward high-visibility dialects at the expense of less-represented varieties. **Practical strategy:** At each release, sample outputs and compare dialect marker frequencies against baseline corpora. Pay particular attention to high-risk categories: function words (where MSA alternatives are common), verbal morphology (where dialect-specific conjugation patterns may be replaced by MSA forms), and discourse markers (which are highly dialect-specific). Prior work on multi-dialectal stance detection [Charfi et al., 2024b] demonstrates that annotation frameworks can effectively track cross-dialect variation, providing methodological models for drift auditing.

6.4. Generalizing Beyond Arabic

The Arabic case illustrates patterns that generalize to other language situations. *Diglossic languages* such as Swiss German, Haitian Creole, and Tamil exhibit similar high-low variety configurations requiring similar attention to prestige drift. *Pluricentric languages* with multiple standard varieties (German, Portuguese, Spanish) face analogous questions about which variety receives implicit priority in LLM outputs. *Languages with limited written tradition*, including many indigenous and minority languages, lack established writing systems; the orthographic preservation principle (Guideline 3) applies through documentation of transcription conventions. *Politically sensitive contexts* exist wherever communities face discrimination, persecution, or cultural appropriation risks; the data sovereignty principle (Guideline 9) generalizes accordingly.

7. Discussion

7.1. Relationship to Existing Frameworks

The proposed guidelines complement existing documentation frameworks. Datasheets for datasets

[Gebru et al., 2021] provide templates for recording corpus composition and intended use. Data statements [Bender and Friedman, 2018] emphasize speaker demographics and situational context. Model cards [Mitchell et al., 2019] document model characteristics and limitations. Our guidelines add dialect-specific requirements that existing frameworks do not fully address: orthographic layering, prestige drift monitoring, and synthetic recursion avoidance.

For community governance and data sovereignty (Guidelines 8–9), our framework aligns with the CARE Principles for Indigenous Data Governance [Carroll et al., 2020], which emphasize Collective benefit, Authority to control, Responsibility, and Ethics. While CARE was developed for indigenous communities, its principles translate well to dialect communities facing similar concerns about data extraction, misrepresentation, and loss of control over linguistic heritage [Rainie et al., 2019].

7.2. Implementation Challenges and Scalability

Implementing the guidelines raises practical challenges. Participatory governance (Guideline 8) requires identifying community representatives, which may be difficult for dispersed dialect populations. Diversity monitoring (Guideline 7) requires defining quantitative metrics for dialect marker presence, which presupposes linguistic analysis of diagnostic features. Long-term stewardship (Guideline 12) requires institutional commitment and sustained funding. We view these challenges as tractable rather than fundamental. The guidelines provide a framework for identifying what needs to be done; implementation pathways will vary by language, community, and institutional context.

A related concern about participatory and human-centered approaches is that they limit scalability. We argue that the guidelines support scalable practices: retrieval-augmented generation (Guideline 2) scales better than unconstrained prompting because it reduces review burden; automated drift monitoring (Guideline 11) enables early detection without exhaustive manual review; pipeline documentation (Guideline 10) enables knowledge transfer across projects. The critical constraint is that scalability must not come at the cost of distributional fidelity.

8. Conclusion

Large language models can support dialect resource creation when they are embedded in archival-first infrastructures, constrained by retrieval, governed by transparent documentation, and validated by dialect communities. Without such safeguards, the same systems can accelerate orthographic homogenization, prestige drift, and synthetic recursion that reduces linguistic diversity.

This paper contributes a theoretical framework, the generator-eraser paradox, and 12 community guidelines for responsible LLM-assisted dialect resource creation. The Arabic case study demonstrates application to a language situation characterized by diglossia and orthographic variability, while highlighting patterns that generalize across language families.

We invite the research community to adopt, adapt, and critique these guidelines, and to develop shared resources that make responsible LLM-assisted dialect resource creation accessible to all communities.

9. Limitations

This paper does not present new experiments and cannot quantify the magnitude of risks for specific models or dialects. The guidelines are intentionally conservative, prioritizing archival integrity and community governance. We have not conducted user studies with resource builders to evaluate usability. The Arabic case study, while detailed, may not capture all relevant phenomena in other diglossic or pluricentric language situations.

While we provide concrete metrics and procedures for diversity monitoring and drift audits (Guidelines 7, 11), we acknowledge that empirical validation would strengthen these recommendations. Future work should include: (1) pilot studies comparing prestige drift rates with and without dialect-aware prompting; (2) longitudinal tracking of dialect marker distributions across staged resource versions using the metrics we propose; (3) controlled comparisons of RAG-grounded versus unconstrained generation for dialect resource tasks; and (4) user studies evaluating the usability of pipeline cards and review rubrics by resource builders with varying technical backgrounds.

We also note that extending these guidelines to multimodal settings (speech, vision-language) raises additional challenges not fully addressed here, including interaction between orthographic variation and ASR/OCR errors, and the need for audio-level provenance tracking.

Future work should evaluate guideline adoption across diverse language contexts, develop shared checklists and templates as community resources, and explore how different dialect communities prioritize preservation versus standardization in applied settings.

10. Ethical Considerations

This work synthesizes existing research and does not collect new data. The guidelines are designed to reduce representational harm and support community sovereignty over dialect resources. We recommend tiered access, opt-out mechanisms, and participatory governance where dialect communities face safety, privacy, or cultural sensitivity risks.

We acknowledge that adopting these guidelines requires resources that may not be available to all teams, potentially creating barriers for under-resourced communities. We encourage well-resourced projects to share infrastructure, documentation templates, and lessons learned to reduce adoption barriers.

Acknowledgments

This work was made possible by the National Priorities Research Program (NPRP) grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), a member of the Qatar Research, Development and Innovation Council (QRDI).

11. Bibliographical References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. NADI 2021: The second nuanced arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021.

El Said Badawi. *Mustawayāt al-ʿArabiyya al-Muʿāṣira fī Miṣr*. Dār al-Maʿārif, 1973.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Douglas Biber. Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4): 243–257, 1993.

Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, 2020.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, 2016.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, 2020. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. The MADAR arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43, 2020.

J.K. Chambers. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Wiley-Blackwell, 2013.

Anis Charfi, Wajdi Zaghouani, S. Hassan Mehdi, and Esraa Mohamed. A fine-grained annotated multi-dialectal Arabic corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2019.

Anis Charfi, Mabrouka Bessghaier, Raghda Akasheh, Andria Atalla, and Wajdi Zaghouani. Hate speech detection with ADHAR: A multi-dialectal hate speech corpus in Arabic. *Frontiers in Artificial Intelligence*, 7:1391472, 2024a.

Anis Charfi, Mabrouka Bessghaier, Andria Sarah Reem Atalla, Raghda Akasheh, Sara Al-Emadi, and Wajdi Zaghouani. MARASTA: A multi-dialectal Arabic cross-domain stance corpus. In *Proceedings of LREC-COLING 2024*, pages 11060–11069, 2024b.

Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, 2020.

Penelope Eckert. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41:87–100, 2012.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

Charles A. Ferguson. Diglossia. *Word*, 15(2):325–340, 1959.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Nizar Habash. *Introduction to Arabic Natural Language Processing*. Morgan and Claypool, 2010.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. A conventional orthography for dialectal arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 711–718, 2012.

Nizar Habash, Salam Khalifa, Fadhl Eryani, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

Clive Holes. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press, 2004.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 2015.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. Gumar: A large scale corpus of gulf arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2016.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Tober, Nana Raffa, Nandini Chadha, Khushbu Das, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117:7684–7689, 2020.

William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, 1972.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Linguistic Data Consortium. Bolt egyptian arabic treebank: Sms and chat. https://catalog.ldc.upenn.edu/LDC2021T17, 2021. Accessed 2026-02-28.

Tony McEnery and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

Ossama Obeid, Go Inoue, Bashar Alhafni, Salam Khalifa, and Nizar Habash. CAMeL tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, 2020.

QCRI. QADI: QCRI arabic dialects identification corpus. https://alt.qcri.org/resources/qadi/, 2020. Accessed 2026-02-28.

Stephanie Carroll Rainie, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodríguez, Jennifer Walker, and Per Axelsson. Issues in open data: Indigenous data sovereignty. In Tim Davies, Stephen B. Walker, Mor Rubinstein, and Fernando Perini, editors, *The State of Open Data: Histories and Horizons*, pages 300–319. African Minds and International Development Research Centre, Cape Town and Ottawa, 2019.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, 2019.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. It's morphin' time! combating linguistic discrimination with inflectional perturbations. In *Proceedings of ACL*, 2020.

Kees Versteegh. *The Arabic Language*. Edinburgh University Press, 2014.

Wajdi Zaghouani. Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Processing Tools, LREC 2014*, 2014.

Wajdi Zaghouani and Anis Charfi. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

Wajdi Zaghouani, Hamdy Mubarak, and Md Rafiul Biswas. So hateful! building a multi-label hate speech annotated Arabic dataset. In *Proceedings of LREC-COLING 2024*, pages 15044–15055, 2024.

Omar F. Zaidan and Chris Callison-Burch. The arabic online commentary dataset: An annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, 2011.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, et al. Findings of the VarDial evaluation campaign. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, 2017.

The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact

Thomas Schmidt[1], **Margaret Blevins**[2], **Hans C. Boas**[2], **Glenn Gilbert**[3]

Department of Germanic Studies, Texas German Dialect Project,
2505 University Avenue, Burdine 336, University of Texas at Austin, Austin, TX 78712-0304
[1]linguisticbits.de, [2]University of Texas at Austin, [3]University of Southern Illinois

thomas@linguisticbits.de, mblevins@utexas.edu, hcb@mail.utexas.edu.
glenngilbert1869@gmail.com

Abstract

The Texas German Dialect Project (TGDP) is a long-standing effort to document the unique variety of German spoken in Texas since the 1840s. For 25 years, the TGDP has built up the freely accessible Texas German Dialect Archive Online (TGDA Online) with recordings and annotations of interviews and language tasks conducted between 2001 and today with some of the last speakers of the variety, which is expected to go extinct within the next 5-10 years. The present paper reports on the most recent addition to the TGDP's online corpus platform— a collection of Texas German data recorded in the 1960s—as well as historical translation elicitations that will be released later in 2026. Both the contemporary and the historical materials follow very similar elicitation methods and are processed using the same pipelines, increasing their comparability. This provides a comparable historical dimension to the resource, enabling diachronic and multidimensional analyses of this endangered variety. These data can help shed light on the dynamics of language contact, dialect contact, and language death.

Keywords: spoken language, variation, corpus, contact variety, Texas German

1. Introduction

Texas German is a unique contact variety for the following reasons: (1) It is a mix of at least five different regional dialects brought by German-speaking immigrants from Europe to central Texas beginning in the 1830s (Boas, 2009). As such, it differs from most other German contact varieties such as Michigan German (Born, 1994), Namibia German (Zimmer et al., 2020), Brazilian German (Altenhofen, (submitted), Kürschner et al., (submitted)), and Hungarian German (Földes, 2002), which can be directly traced back to specific donor dialects from a single region (Boas, 2016). (2) Texas German is, in comparison to other German contact varieties such as Pennsylvania German, Romanian German, and Hungarian German, a relatively young contact variety. (3) Unlike other new world contact varieties of German, Texas German did not evolve into a focused new world variety according to the model of Trudgill (2004). Because intergenerational transmission was severely restricted following World War I, Texas German never went through all three stages of new dialect formation, being "stuck" halfway through the second stage (Boas, 2009).[1] Texas German thus exhibits a very high degree of inter-speaker and intra-speaker variation, making it challenging to provide a clear definition of how Texas German should be characterized linguistically. (4) Texas German is one of the most extensively documented extraterritorial contact varieties of German, making it a particularly fruitful basis for linguistic research.

2. Background

2.1 Similar varieties

Due to research, funding, and time limitations, as well as the availability of the data itself, many corpora of low-resource language varieties contain a potentially limited snapshot of a particular variety (e.g., recordings of natural conversation and elicited translations from a particular 5-year timespan). Or data may be collected as part of a particular research project and then not be made accessible or public. It can therefore be challenging to find accessible data that span several timeframes, genres, and elicitations methods.

There are numerous corpora of (Germanic) heritage languages, all of which have different coverages, timeframes, and accessibility. For example, there are recordings of Wisconsin German (USA) from the 1960s and 2001 (Eichhoff 1979, Wagener 2002), but only the former is currently publicly available. The Corpus of American Danish (CoAmDa) contains recordings from both North America (three sets: 1960s-1980s, early 1990s, and late 1990s) and Argentina (2014-2015). CoAmDa is fully transcribed and marked for language (Kühl et al. 2017). The MEND corpus, a collection of Mennonite German (Kaufmann et al. 2023), contains a large set of translation elicitation

[1] Trudgill's three stages of dialect formation are (1) (first generation): rudimentary leveling, (2) (second generation): variability and mixing, (3) (third generation): emergence of stable, relatively uniform dialect.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 221–229
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

recordings from five different countries from a single time period (1999-2002). A similar situation is true for the RUEG corpus, which is a highly multifaceted collection containing several kinds of data (written and spoken, informal and formal) from a variety of (heritage) language speakers (mono- and multi-lingual speakers of Greek, Russian, German, and Turkish, English, and Kurdish) from multiple locations (Greece, Russia, Germany, Turkey, and the USA) during a ~6 year timespan (Schroeder et al. 2024).[2]

For Texas German, there is a plethora of data available covering many time periods, modes, and genres, including:

- Narrative audio interviews from the 1960s, 1990s, and 2001-today (~500 hours)
- Audio recorded translation tasks from the 1960s, 1990s, and 2001-today (~350 hours)
- Letters and diaries from the mid-1800s through the early 1900s (2,000+ pages)
- Newspapers from the mid-1800s through the mid-1900s (800,000+ pages)
- Reports from the mid-1800s through the early 1900s (1,000+ pages)
- Novels and poetry from the late-1800s through the 1900s

This puts researchers in the unique position to be able to explore the evolution of a heritage contact variety through over 100 years of development from multiple angles.

In what follows, two comparable sets of data are described. Section 2.2 discusses a set of audio recordings of contemporary Texas German (2001-today) and section 3 introduces a set of Texas German recordings from the 1960s. Both sets of data were elicited in similar ways as well as processed and annotated using the same or very similar methods with the goal of making a resource that facilitates diachronic research.

2.2 The TGDP and the TGDA Online

Since 2001, members of the Texas German Dialect Project (TGDP) have interviewed more than 800 speakers of Texas German, recording three different types of data. The first type of data consists of participants' translations of English word lists and sentences taken from Eikel (1954), Gilbert (1972), and Guion (1996) into Texas German. The second type of data consists of sociolinguistic, open-ended interviews, which aim to capture the informants' daily use of Texas German. These interviews seek to elicit casual, relaxed conversations in which speakers produce Texas German spontaneously without being asked to produce specific linguistic structures. The third type of data is a biographical questionnaire that captures relevant extralinguistic information about each speaker, including age, religious affiliation, language use throughout the lifespan, and language, and speaker attitudes.[3]

Audio recordings of participants' translations based on Eikel (1954) and Gilbert (1972)'s wordlists are available at https://tgdp.org/. Data for 750+ speakers from 90+ locations is currently available online.

Audio recordings of open-ended interviews, along with their time-aligned transcripts, are available at https://tgdp-zumult.la.utexas.edu/.[4] The most recent December 2025 release of the data contains 196 hours of audio recordings (1,494,400 tokens, 475 speakers).[5] The transcripts available contain a literary transcription, a tokenized layer, language tags, normalization, part-of-speech tagging (one layer according to STTS 2.0 (Westpfahl & Schmidt, 2016) and another layer for Universal Dependencies POS (Nivre et al. 2020)), lemmatization, a canonical phonetic transcription based on the transcribed forms, and a word-for-word English translation. These transcripts, along with all the annotation layers, can be queried using a CQP query tool.

3. Integrating Gilbert's Historical Recordings

3.1 Background

In the 1960s, Glenn Gilbert recorded several speakers as part of the fieldwork for his dissertation *The German Dialect Spoken in Kendall and Gillespie Counties, Texas* (Gilbert, 1963). Several years later, he interviewed dozens more Texas German speakers, mostly hand-transcribing the interviews in situ, but also audio recording some of them. This fieldwork, in addition to the fieldwork he had already done for his dissertation, laid the groundwork for his *Linguistic Atlas of Texas German* (Gilbert, 1972). In 2017, Gilbert donated all his audio recordings, fieldnotes, and several boxes of other research materials related to his research on Texas German to the TGDP.

These audio recordings create a unique situation: With access to audio recordings from the 1960s and from 2001 to today, it is possible to make

[2] https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/rueg/rueg-corpus

[3] For details on the interview process, the processing of recordings, the workflow for transcribing and translating the interviews, and how the data are stored, see Boas (2021) and Boas et al. (in press).

[4] This corpus is available via an instance of the ZuMult platform (Fandrych et al., 2023; Boas et al., in press).

[5] Because interviews, audio processing, and transcription are ongoing, only a portion of the interviews recorded by the TGDP are currently available online (~60% of the open-ended interviews, ~90% of the translation interviews).

comparisons between the two sets of data, allowing for a better understanding of how Texas German has evolved over the last 50+ years.[6] This can also help us better understand how languages in contact may evolve over time in general.

3.2 Available Materials

The audio recordings Gilbert donated to the TGDP include sub-collections of other contact varieties, such as Texas Czech, Texas English, and Kansas German. The Texas German portion of his audio collection is referred to as the GTXG corpus ('Gilbert Texas German').[7]

GTXG contains 63 *events* with 98 participants.[8] The term "event" refers to the entirety of recordings made in one informant session (sometimes with one speaker, sometimes with a group of speakers). Each event consists of one or more *speech events*. The term "speech event" refers to one particular type of recorded speech (e.g., translation elicitations, free conversation, etc.). The speech event types are as follows:

- Open-Ended (OE): sociolinguistic, narrative, open-ended interviews in which a non-Texas German speaker talks with a Texas German speaker about their life
- Atlas (GA): A translation elicitation interview in which the interviewer asks the participant to translate a set of English words and phrases into Texas German. This list (~230 prompts) is most similar to the set of words and phrases used to construct Gilbert's *Linguistic Atlas of Texas German* (1972)
- Long (GL): A translation elicitation interview in which the interviewer asks the participant to translate a set of over 1,200 English words and phrases Texas German.[9]
- Wenker sentences (WS): The list of sentences Georg Wenker used to conduct a survey of Germany dialects between 1876 to 1887 (~115 prompts)[10]
- Less frequent translation lists (English into Texas German):
    - Miscellaneous (GM): ~150 prompts
    - Short (GS): ~100 prompts[11]
    - Extra (GE): ~300 prompts
- Biographical interview (BI): Participants' responses to a set of biographical questions, such as when and where they were born.

The amount of data for these eight speech event types is depicted in Table 1 below.

The lengths of the translation lists vary considerably. For example, the "Gilbert Long" (GL) translation list has over 1,200 prompts, while "Gilbert Short" (GS) has a little over 100 prompts. Some translation prompts appear in multiple translation lists.

Interview Type		Amount
Open-Ended (OE)		~ 51 interviews, ~ 11 hours
Translation	Atlas (GA)	~ 28 interviews, ~ 7 hours
	Long (GL)	~ 12 interviews, ~ 21 hours
	Wenker sentences (WS)	~ 16 interviews, ~ 2 hours
	Other (GM, GS, GE)	~ 9 interviews, ~ 1.5 hours
Biographical Interview (BI)		~ 49 interviews, ~ 8 hours

Table 1: GTXG Speech Events

3.3 Data Processing

3.3.1 Processing of Open-Ended Interviews

The GTXG open-ended interviews are segmented into smaller thematic chunks, manually transcribed and translated using the transcription software ELAN[12] and a tokenized layer, language tags, normalization, two layers of part-of-speech tagging (STTS 2.0 and Universal Dependencies POS), lemmatization, and a canonical phonetic transcription based on the transcribed forms are added automatically.[13] Almost all the GTXG open-ended interviews have been transcribed and are available via the ZuMult corpus platform at https://tgdp-zumult.la.utexas.edu/.[14]

[6] There is also one person who was recorded by both Gilbert and the TGDP, allowing for a case study of how a single person's language has changed over time.

[7] Please note that GTXG is referring to the Texas German recordings that Gilbert made in the 1960s and 1970s. In Table 1, "Gilbert" is referring to the collection of recordings the TGDP had made using Gilbert's elicitation list, but these recordings were all made after 2001.

[8] About a quarter of these participants are incidental speakers, i.e., people who were in the room when the primary participant was being interviewed, but who did not speak much themselves.

[9] See Gilbert (1963: 29-63).

[10] See Fleischer (2017) for more information about the Wenker sentences and surveys.

[11] See Gilbert (1963: 78-81).

[12] ELAN (https://archive.mpi.nl/tla/elan) at the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands (see Brugman, and Russel 2004).

[13] These automatically added annotations are the same layers as are present in the TGDP Corpus annotations of contemporary Texas German and are created using the same pipeline.

[14] The only files that have not been released are the audio files from 2 events (1 in Wendish, 1 in Alsatian), as these files could not be properly quality checked yet.

3.3.2 Processing of Translation Interviews

When building spoken language corpora, focus is often put on more naturalistic speech. Within dialectology studies, however, elicitation lists are also commonly used, but they are often not made searchable in the same way more naturalistic spoken language often is. To make the speech events that were based on elicitation lists more accessible to researchers, the TGDP has developed a system to transcribe them. This transcription system involves several annotation tiers, facilitating different ways to explore the data.

The TGDP is transcribing the elicitation list speech events in phases, starting with the "Gilbert Atlas" speech events (GA). To do this, GA speech events were first segmented using the audio editing software Audacity (Audacity Team 2014). During this phase, students were asked to add labels to all the portions of the audio during which a speaker was providing an answer to a translation elicitation and label that portion of the audio with the appropriate elicitation number. If a student noticed that the interviewer asked for an elicitation which was not on the original list, they were instructed to add it to the list and give it a new number. This was to ensure that the prompt list accurately depicted what the informants were being asked to translate.

Following this step, the translation interviews were processed in two separate ways in parallel. To test the potential of Automatic Speech Recognition (ASR) for economizing on transcription time, the audio files were first processed with Whisper (Radford et al., 2023). For each section marked in Audacity, a suitable prompt of the following kind was generated and sent to the OpenAI platform alongside the corresponding audio.

Hier ist ein Satz in englischer Sprache: „This chicken has long feathers". Können Sie bitte diesen Satz in die deutsche Sprache übersetzen?
[Eng: Here's a sentence in English language: … Can you please translate this sentence into the German language?]

The expectation was that by providing such a highly restricting prompt, useful results could be achieved in spite of the suboptimal audio quality and the non-standard language used by the speakers. At first glance, results looked usable but nowhere near error-free. Students were asked to correct a sample of the results, confirming the findings in Gorisch and Schmidt (2024) that there is hardly a gain in efficiency over fully manual transcription with this type of ASR output. In other words, it took students the same amount of time to transcribe the audio from scratch as to correct a Whisper transcription, and the transcripts that the students constructed tended to keep some of Whisper's normalizations that the TGDP did not want in its transcripts (e.g., standardizing non-standard language, skipping over disfluencies, misunderstanding words, etc.).

Students were therefore asked to transcribe the GA speech events from scratch, focusing only on the portions of the audio that captured participants' responses.[15] These transcriptions were created in ELAN and had four kinds of layers: (1) a transcription tier for each speaker, (2) a prompt tier (PROMPT), (3) a prompt number tier (NO), and (4) a target hypothesis tier (TH). To avoid bias, students were not given the Whisper-generated transcriptions. This resulted in files that looked like Figure 1 below.

In the transcription tier, speakers' utterances were transcribed using modified German orthography to try to capture participant's actual speech with deviations from the norm. The target hypothesis (TH) tier provides an interpretation of the actual speech into the translation the speaker "intended" without any disfluencies or interactive elements of the performed speech. For example, if a speaker said "Uhhm … Geschirr- Geschirrschüssel" as a response to the question "How do you say 'sink' in Texas German?", then the Target Hypothesis would simply be "Geschirrschüssel", i.e., the participants complete "final answer".

Several special characters were used to mark the following special situations (see Table 3 below for a summary). If a participant produced multiple answers, all their answers are listed in the TH layer, separated by a pipe (|). If a word or set of words is "offered" by the interviewer or another speaker and then "accepted" by the participant, each of the "accepted" words are marked with a %. Any words that are implied by the context but not explicitly said by the speaker are marked with a plus sign (+). If the speaker does not provide a translation, then /NO_ANSWER/ is written in the TH tier. Making these distinctions anticipates different ways of analysing the data where, for instance, suggested words will have to be ignored in one approach but will be relevant in another.

[15] This decision was made in an effort to capture the most core data while reducing transcription time. In a future phase, the interviewer's audio could be transcribed as well. This would be helpful for studies interested in e.g., what participants struggle with (for example, by looking at how long it takes a participant to start answering once a prompt is given).

The special symbols can be transferred to explicit annotations in later processing.

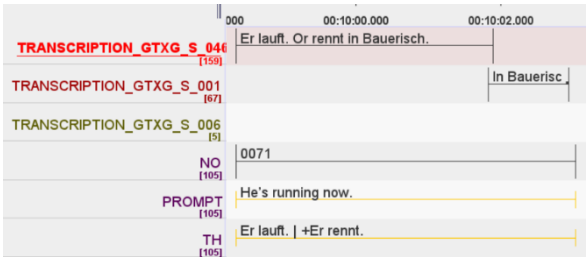


Figure 1: Screenshot of Excerpt from GTXG_E_024_SE_GA Transcript in ELAN.

1. *Er lauft. Or rennt in Bauerisch.*
   he runs or runs in Bauerish
   'He runs. Or 'runs' in Bauerish (a language variety)'

The example in Figure 1 above captures two of these situations. First, the participant offered multiple translations, both of which are captured in the TH layer, separated by a pipe (|). Second, the "er" in "+Er rennt." is marked with a "+" because the *er* 'he' is the implied subject based on the participant's first translation. The speaker does not actually say *er rennt*—he only implies that *er* would be part of his second response.[16]

Target Hypothesis Symbol	Meaning
|	The participant offered multiple translations, and this symbol separates each of them.
%	Someone other than the participant "offered" this token to the participant, and the participant accepted it.
+	The participant did not literally say this token, but it is implied by the surrounding context.
/NO_ANSWER/	The participant does not provide a translation for this prompt.

Table 2: GTXG Target Hypothesis Symbols

As mentioned above, the TGDP is beginning their transcription of elicitation speech events with the "Gilbert Atlas" speech events (GA). The GA portion of the GTXG corpus includes 32 Texas German speakers across 22 events and a list of ~230 translation prompts.[17] There is a total of ~25,000 transcribed tokens and 2,469 recorded target hypotheses.[18] The occurrence of the TH special symbols is summarized in Table 3 below, which can be read as follows: The pipe symbol "|" occurred a total of 68 times in the GA portion of the GTXG corpus. These occurrences spanned 50 different elicitation questions. In other words, there were some elicitation numbers for which multiple speakers provided multiple translations, but generally speaking, participants provided multiple translations for different elicitation prompts. The 68 tokens of the "|" symbol were produced by a total of 20 speakers, with some participants only offering multiple translations once, while other participants produced multiple translations for 6-8 prompts.

This kind of information can aid research on participants' confidence and flexibility. Information about what participants do *not* answer can help shed light on which lexical items and syntactic structures may be more challenging for participants.

Symbol	Tokens	Elicitations	Speakers
|	68	50	20
%	309	80	26
+	193	39	25
/NO_ANSWER/	118	47	24

Table 3: Target Hypothesis Symbols in the GA Portion of the GTXG corpus

3.4 Metadata

The recordings that Gilbert donated to the TGDP did not come with consistent metadata. Therefore, much of the interview/speaker metadata must be gleaned from the audio recordings themselves, as well as relevant publications. There is basic metadata available for ~20 participants (based on Gilbert 1972) and more comprehensive metadata for 7 participants (~30 questions, Gilbert 1963).

4. Example Queries

The availability of (transcribed) Texas German free conversation and translation elicitation audio from both the 1960s and from the last 25 years allows for various kinds of comparative research. On the one hand, one can use the open-ended data to explore how conversational language may

[16] The annotation schema described in this paper for translation elicitations is a novel annotation schema. Guidelines with more examples will be made available when this dataset is made public later this year.

[17] Originally, there were about ~150 prompts in Gilbert's prompt list, including some built in repetitions of prompts. In the process of identifying the individual responses for transcription purposes, another ~80 prompts were identified, often in the form of clarification questions, rephrases of prompts (e.g., 'the cow' vs. 'what do you call the animal that you get milk from'), or small deviations from the original prompts.

[18] Note: In the current iteration of the transcripts, the interviewer's prompt is not transcribed.

have changed within the last 50+ years. On the other hand, one can compare the results from different elicitation methods (conversational data vs. elicited data) to see whether and how results may vary based on elicitation type. Brief examples of both kinds of research are provided below (see also screenshots in the appendix).

4.1 Diachronic Comparison

One way of using the available Texas German audio recordings is to compare how language has (or has not) changed since the 1960s. For example, if one were interested in looking at whether /ɹ/ (a voiced alveolar approximant) has increased in usage since the 1960s, one could search all normalized tokens that contain <r> and listen to their phonetic realization (CQP: `[norm=".*r.*"]`). This could be of interest because /ɹ/ was not present in the German phonological inventory of the German dialects that the majority of German-speaking immigrants who came to Texas spoke, but it is prevalent in American English, and thus could indicate an increasingly Americanized pronunciation of German lexemes due to prolonged contact with English.

Another aspect of language change that could be explored would be to look at the percentage of English loanwords present in Texas German speech. On a basic level, one could conduct the query `[lang="deu"][lang="eng"][lang="deu"]` to find all tokens tagged as English that are surrounded by a token tagged as German on either side.[19] This search finds 11,237 hits in the contemporary TGDP data (i.e., 0.75% of the 1,494,400 total tokens) and 28 tokens in the GTXG corpus (i.e., 0.03% of the 96,309 tokens). Although a more in depth look at the data is necessary, this could indicate that (a) there is a relatively low number of English loanwords and/or nonce borrowings in Texas German, despite it being in contact with English for over 150 years and (b) the number of individual English loanwords in Texas German has increased over the last 50 years.

4.2 Free Conversation vs. Translation Elicitation

Translation elicitations can be a helpful tool in that they ensure that a participant is asked to produce a particular set of lexemes/features/constructions that may not have (frequently) occurred in a less structured conversation.[20] They can also ensure that a set of participants are all asked the same questions, making it easier to compare participants' responses. It is also possible, however, that the prompt itself influences how a participant produces a particular utterance (cf. Bailey et al. 1997, Munro 2022, Vintoniak et al. 2024). In order to get a more thorough picture of a particular variety, it can therefore be helpful to look at data that was collected using multiple elicitation strategies.[21]

For example, one could look at all the translations of prompts "There is the man who I want to see" and "There are the children who I gave the candy to" to get one view of how Texas Germans produce relative pronouns (i.e., how they choose to translate 'who'). Then, one could search for all the relative pronouns in the conversational transcripts and see if speakers used the same strategies within the same contexts (e.g., animate vs. inanimate referents).

5. Conclusion and Outlook

We have introduced the spoken language resources compiled by the Texas German Dialect Project, particularly highlighting the new addition of historical recordings from the 1960s. These data, in addition to the contemporary data from the last 25 years, can facilitate a diachronic exploration of Texas German and support comparative studies.

The open-ended interviews from the historical data (GTXG) were published to the ZuMult platform in December 2025. In 2026, the first phase of the historical elicited translation data will be added (GA speech events within GTXG), and the platform will be extended with functionality to make optimal use of the comparability inherent in this part of the data. In the mid-term, the TGDP will begin processing the corresponding translation data from the (much larger) "modern" corpus in a similar manner with the goal of making it available in ZuMult. In this context, we will evaluate the ASR results in more detail by comparing them to the manual transcriptions currently being created. This should give us a clearer idea if and how ASR can be made fruitful for this type of data.

The historic dimension in the data is currently already made use of in the project "Let the People of the Past Speak! Turning Migrant Letters of the 19th Century into Speech", which aims to reconstruct the language and speech of German immigrants in the 19th century. Comparing the Texas German data from the 21st century to the older recordings is used as a basis to extrapolate

[19] How the search tool is currently configured, queries automatically only search participant transcriptions (i.e., do not include interviewer utterances). Interviewer utterances can, however, be included in searches.

[20] There are many kinds of (linguistic) elicitation strategies, see e.g., Arunachalam (2013), Chambers and Trudgill (2004), and Kainada and Baltazani (2013).

[21] Lenz (2016) is one of many researchers who suggests it is best to use a variety of elicitation methods when investigating a particular set of linguistic phenomena.

certain developments (e.g., dialect levelling, language mixing) to an earlier stage.

6. Availability

The TGDP Corpus data, including the GTXG data, that are available via https://tgdp.org/ and https://tgdp-zumult.la.utexas.edu/ are available for research and teaching purposes. Use of the data for commercial purposes or training of AI requires additional agreements to be negotiated with the TGDP director. More information about user rights and responsibilities are listed at https://tgdp.org/dialect-archive/.

7. Acknowledgments

Over the years, the Texas German Dialect project has profited from financial support from the University of Texas at Austin, the National Endowment for the Humanities, the Texas Council for the Humanities, the Consul General of the Federal Republic of Germany in Houston, Texas, the Alexander von Humboldt Foundation, the Raymond Dickson, Alton C. Allen, and Dillon Anderson Centennial Professorship, the Alkek Foundation, the Texas German Dialect Project Endowment, the Dan L Duncan Foundation, and many private donors who wish to remain anonymous. The project "Let the people of the past speak" is funded by the Volkswagen foundation under grant numbers 9E083 and 9E529. The corpus platforms for the TGDP data were constructed in collaboration with the Liberal Arts Instructional Technology Services at the University of Texas at Austin and linguisticbits.de. We would like to thank all TGDP members and student assistants for their contributions to the project, and, especially, the speakers of Texas German for their willingness to provide their data.

8. Ethics Statement

All participants were informed about the project and the intended use of the recordings. Written consent was obtained for recording, transcription, and inclusion of their data in the corpus. Any directly personally identifying information was removed during transcription and in the audio. When making data available, we are complying with FAIR principles, observing the restrictions resulting from the informed consent. We are also respecting CARE principles by allowing participants to decide whether their recordings are made publicly accessible, and by giving participants the option of having copies of their unedited recordings for their own personal use.

9. Bibliographical References

Altenhofen, C. (submitted). Hunsrückisch in Brazil: The dynamics of a German variety outside of the "Heimat". In Boas, H., Blevins, M., P. Wolf-Farré (Eds.), *German language islands in Latin America*.

Arunachalam, Sudha (2013). Experimental Methods for Linguistics. *Language and Linguistics Compass* 7(4): 221–232, 10.1111/lnc3.12021 Audacity Team (2021).

Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved March 17th 2021 from https://audacityteam.org/

Bailey, G., Wikle, T., and Tillery, J. (1997). The effects of methods of results in dialectology. *English World-Wide* 18(1), pp. 35-63.

Blevins, M. (2022). *The language-tagging & orthographic normalization of spoken mixed-language data, with a focus on Texas German*. Ph.D. thesis. The University of Texas at Austin.

Boas, H. (2009). *The life and death of Texas German*. Duke University Press, Durham.

Boas, H. (2016). Variation im Texasdeutschen: Implikationen für eine vergleichende Sprachinselforschung. In A. Lenz (Ed.), *German Abroad: Perspektiven der Variationslinguistik, Sprachkontakt- und Mehrsprachigkeitsforschung*, 11–44. Vienna University Press.

Boas, H. (2021). Zwei Jahrzehnte digitale Dokumentation und Erforschung eines aussterbenden deutschen Auswanderer-dialekts: Das Texas German Dialect Project (2001-2021). *Zeitschrift für deutschsprachige Kultur und Literatur* 30, pp. 234–268.

Boas, H., Schmidt, T., & Blevins, M. (in press). A new corpus platform for the Texas German Dialect Project. *Language Resources and Evaluation*.

Born, R. (1994). *Michigan German in Frankenmuth: Variation and change in an East Franconian dialect*. Camden House, Columbia, South Carolina.

Brugman, H. & Russel, A. (2004). Annotating multimedia / multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004), pp. 2065–2068, Lisbon, ELRA.

Chambers, J. K., and Trudgill, P. (2004). *Dialectology*. Cambridge University Press, Cambridge.

Eichhoff, J. (1979). Deutsche Sprache in Wisconsin. In L. Auburger, H. Kloss, & H. Rupp (Eds.) *Deutsch als Muttersprache in den Vereinigten Staaten, Teil 1 Der Mittelwesten. Deutsche Sprache in Europa und Übersee. Berichte und Forschungen. Vol. 4*. Franz Steiner Verlag, Wiesbaden, pp. 65–75.

Eikel, F. (1954). *The New Braunfels German dialect*. Manuscript. Johns Hopkins University.

ELAN (Version 7.0) [Computer software]. (2025). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan.

Fandrych, C., Schmidt, T., Wallner, F. & Wörner, K. (Eds.) (2023). *Korpora Deutsch als Fremdsprache* 3(1). Themenschwerpunkt:

Zugänge zu mündlichen Korpora für DaF und DaZ: Das ZuMult-Projekt Darmstadt: KorDaF.
Fleischer, J. (2017). Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen: Dokumentation, Entdeckungen und Neubewertungen. *Deutsche Dialektgeographie* 123. Hildesheim/Zürich/New York: Olms.
Földes, C. (2002). Kontaktsprache Deutsch: Das Deutsche im Sprachen- und Kulturenkontakt. In U. Hass-Umkehr, W. Kallmeyer, G. Zifonun (Eds.), *Ansichten der deutsche Sprache. Festschrift für Gerhard Stickel zum 65. Geburtstag*, pp. 347-370. Narr, Tübingen.
Gilbert, G. (1963). *The German dialect spoken in Kendall and Gillespie counties, Texas*. Ph.D. thesis, Harvard University.
Gilbert, G. (1972). *Linguistic atlas of Texas German*. University of Texas Press, Austin.
Gorisch, J. & Schmidt, T. (2024). Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024). ELRA, pp. 6564–6574.
Guion, S. (1996). The death of Texas German in Gillespie County. In P. S. Ureland, I. Clarkson (Eds.), *Language and contact across the North Atlantic*, pp. 443–463. Niemeyer, Tübingen.
Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36, pp. 161–195.
ISO (2016). ISO 24624:2016 Language resource management — Transcription of spoken language.
Kainada, E. and Baltazani, M. (2013). Evaluating methods for eliciting dialectal speech. In *M. Janse, B. D. Joseph, A. Ralli* and *M. Bagriacik* (Eds.), *Proceedings of the 5th International Conference on Modern Greek Dialects and Linguistic Theory, Patras,* pages 101–123.
Kaufmann, G., Gorisch J., Schmidt, T. (2023). Das MEND-Korpus im Archiv für Gesprochenes Deutsch: Entstehung, Möglichkeiten, Grenzen. In P. Wolf-Farré, L. Löff Machado, A. Prediger, & S. Kürschner (Eds.). *Deutsche und weitere germanische Sprachminderheiten in Lateinamerika: Methoden, Grundlagen, Fallstudien*. Peter Lang, Berlin, pages 103–147.
Kühl, K., Heegård Petersen, J., Foget Hansen, G., and Gregersen, F. (2017). CoAmDa: et nyt dansk talesprogskorpus. *Danske Talesprog* 17, pp. 131–60.
Kürschner, S., Habermann, M., Prediger, A., & Pupp Spinassé, K. (submitted). Bohemian varieties in Southern Brazil. In Boas, H., Blevins, M., P. Wolf-Farré (Eds.), *German language islands in Latin America.*
Lenz, A. N. (2016). On eliciting dialect-syntactic data. Comparing direct and indirect methods. In A. Speyer, P. Rauth (Eds.), *Syntax aus Saarbrücker Sicht. Beiträge der SaRDiS-Tagung zur Dialektsyntax* (Zeitschrift für Dialektologie und Linguistik Beihefte. 165), pp. 187-219. Stuttgart: Steiner.
Munro, M. (2022). Variability in L2 vowel production: Different elicitation methods affect individual speakers differently. *Frontiers in Psychology 13.*
Nivre, J. de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4034–4043, Marseille, France.
Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Doi: 10. 48550/arXiv.2212.04356
Schroeder, C., Lüdeling, A., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Grigoriadou, S., Hartz, R. G., Iefremenko, K., Jahns, E., Katsika, K., Keller, M., Klotz, M., Krause, T., Labrenz, A., Martynova, M., Özsoy, O., Pashkova, T., Pohle, M., and Zürn, N. (2024). RUEG Corpus. 10.5281/zenodo.11234583.
Vintoniak, O., Hnatyyuk, M., Miniailo, R., Turysheva, O., and Kotvytska, V. (2024). Dialectology in modern linguistic research: Theoretical approaches and methods. *Journal of Interdisciplinary Research* 14, pp. 39-44.
Wagener, P. (2002). German dialects in real-time change, *Journal of Germanic Linguistics* 14(3): 271–285.
Westpfahl, S. & Schmidt, T. (2016). FOLK-Gold – A GOLD standard for part-of-speech-tagging of spoken German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). ELRA, pp. 1493-1499.
Zimmer, C., Wiese, H., Simon, H., Zappen-Thomson, M., Leugner, J., Bracke, Y., Stuhl, B., & Schmidt, T. (2020). Das Korpus *Deutsch in Namibia* (*DNam*): Eine Ressource für die Kontakt- Variations- und Soziolinguistik. *Deutsche Sprache* 3, pp. 210-232.

10. Language Resource References

The Texas German Dialect Project. 2001+. *The Texas German Dialect Archive Online.* https://tgdp.org/

The Texas German Dialect Project. 2001+. *The Texas German Dialect Project Corpus*. distributed via ZuMult. https://tgdp-zumult.la.utexas.edu/

Gilbert, G., The Texas German Dialect Project. 1962+. *The Gilbert Texas German Corpus* distributed via ZuMult. https://tgdp-zumult.la.utexas.edu/

11. Appendix

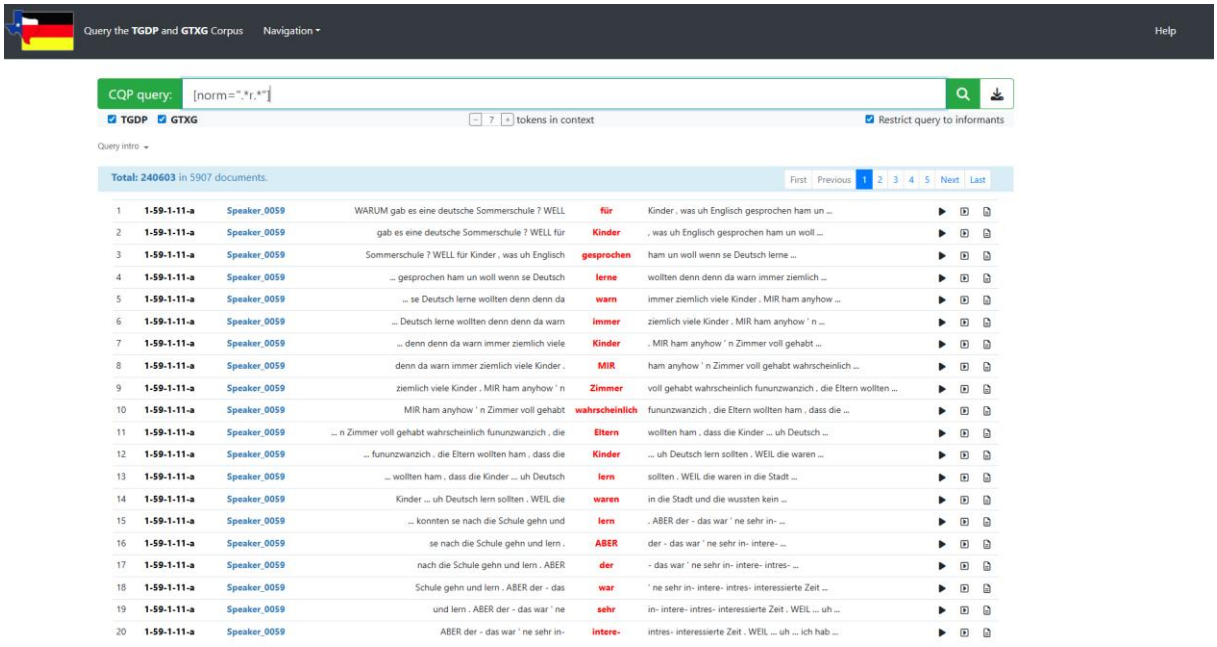


Figure 2: CQP Query `[norm=".*r.*"]` – finding all normalized tokens containing an -r-

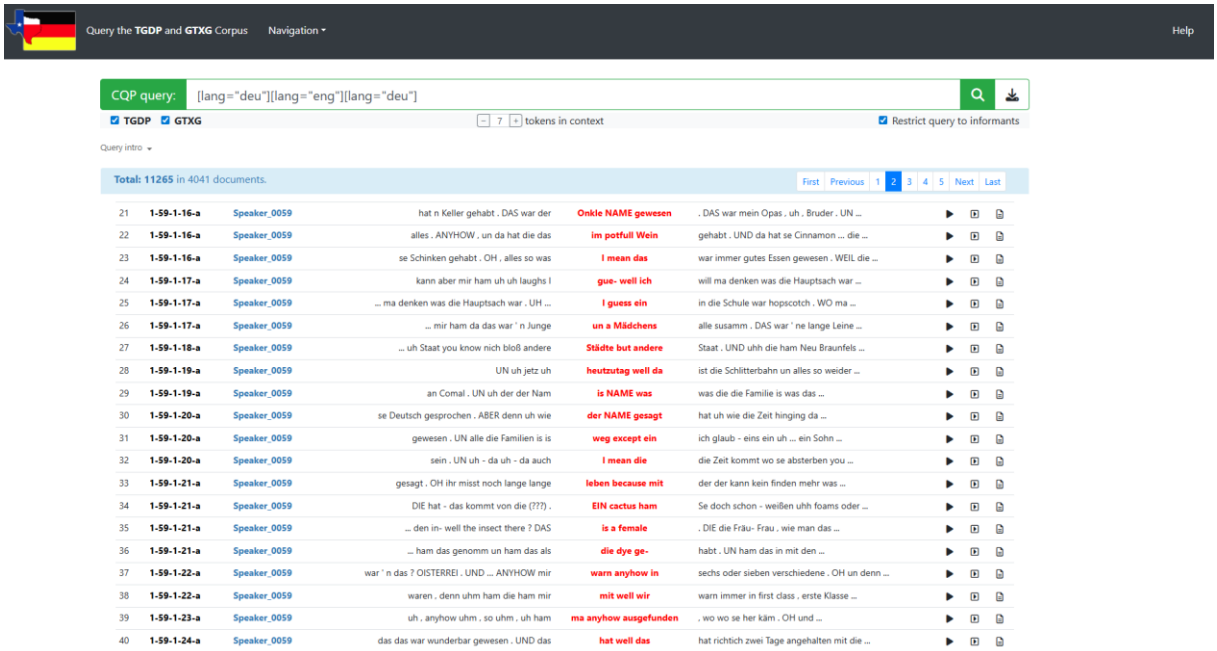


Figure 3: CQP Query `[lang="deu"][lang="eng"][lang="deu"]` – finding all tokens tagged as English that are surrounded by a token tagged as German on either side

Pontic Greek in the Caucasus: an online corpus

Svetlana Berikashvili,[1] **Stavros Skopeteas**[2]
Ilia State University,[1] University of Göttingen[2]
Kakutsa Cholokashvili Ave 3/5, Tbilisi 0162,[1] Kate-Hamburger-Weg 3, Göttingen 37073[2]
svetlana.berikashvili@iliauni.edu.ge,[1] stavros.skopeteas@uni-goettingen.de[2]

Abstract

This article presents a multi-media corpus of Pontic Greek as spoken in the Caucasus (Georgia). The corpus covers three major stages reflecting different sociolinguistic settings: (a) rural communities in the original settlements in Georgia; (b) internal migration to urban centers within Georgia; (c) external migration to Greece. The dataset comprises 373 audio recordings (total duration 7h 26m; total word count: 43,073) and is released as an open-access resource. It includes audio files (WAV) and annotations (XML). Annotations provide orthographical transcription, morphemic transcription and morpheme-by-morpheme and sentence-by-sentence translations in English (Toolbox). Transcriptions are time-aligned with the audio files (ELAN) and can be queried online (ANNIS). The corpus is intended for linguists working on dialectology, language contact, and language change, as well as for a broader audience interested in the history and practices of Pontic Greek communities. Pontic Greek in the Caucasus offers a unique opportunity to investigate contact between Greek and Indo-European (Russian) as well as Non-Indo-European languages (Georgian, Turkish) in different sociolinguistic settings.

Keywords: Pontic Greek, corpus resource, corpus design, corpus annotations, endangered dialects

1. Introduction

The development of dialectal oral and textual resources is essential for our knowledge of linguistic heritage and language variation. This paper presents a dataset designed to capture variation in Pontic Greek, as spoken by the Pontic speakers of Georgia, whose ancestors migrated from Anatolia in several waves beginning in the nineteenth century. Pontic Greek in the Caucasus remains underrepresented in research on Greek dialects and, in light of the migration processes of the past three decades, has become severely endangered. Its systematic documentation in online resources is therefore an urgent priority.

The data for the present corpus were collected from Pontic speakers in Georgia as well as from speakers who have migrated to Greece. The dataset comprises a morphologically annotated corpus of narrative texts and additional interviews (with a total recording time of 7h 26m). Beyond their linguistic relevance, these materials also provide insights into cultural practices. The resource is available both as an online corpus at `https://spw.uni-goettingen.de/projects/xtyp/PNT.html` and as downloadable files through the TLA (Kotanidi et al., 2019).

The relevant background of Greek populations in the area is outlined in Section 2. Section 3 reviews related resources on Greek dialects. Section 4 presents the methodology of the data collection. Section 5 outlines the annotations of the corpus and Section 6 the online resources. Section 7 illustrates the use of the corpus, and Section 8 concludes.

2. Greek Populations in the Caucasus

The ethnic Greek population in the Caucasus consists of two linguistic groups: (a) Urum speakers, whose language is closely related to Anatolian varieties of Turkish, and (b) Pontic speakers, who use the Pontic dialect of Greek; see Sideri (2006) and Höfler (2020) on the interplay between language and ethnic identity. Both groups migrated to Georgia from the Ottoman Empire in several waves. Approximately 800 families had already settled in eastern Georgia in the eighteenth century (Kaukhchishvili, 1946); Greek villages in Armenia were likewise established in the eighteenth century as a result of labor migration (Hodgson, 2010). Major migration waves to Georgia followed the end of the Russo-Turkish war of 1828-1829 (42,000 people), the Crimean war 1853-1856 (36,000 people), and a further wave took place between 1878 and 1882 (17,100 people) (Xanthopoulou-Kyriakou, 1997).

Pontic Greeks originated from various areas of northeastern Anatolia, including Kars, Erzurum, Trabzon, and Gümüşhane (see triangles in Figure 1). They settled in numerous villages and towns across Georgia. The earliest documented settlement (since 1829) is Ts'ints'q'aro (Kaukhchishvili, 1942). Other villages are located in Tetritskaro region (e.g., Big Iraga, Small Iraga, Jigrasheni, etc.), in Dmanisi (Gora, Sakire), in Borjomi (Tsikhisdjvari), along the Black Sea coast in Adjara (Dagva, Kvirike, Akhalsheni) as well as in Abkhazia (Aleksandrovka, Mikhailovka). Most villages in the Ts'alka region (south of Tbilisi) are Urum-speaking, although Pontic Greek is also reported in some settlements

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 230–237
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

(Santa, Neokharaba). In northern Armenia, the Greek villages are concentrated around Hankavan.

Official records only report ethnic affiliation, but do not distinguish groups by language. The ethnic Greek population in Georgia reached its peak in 1989, with 100,324 individuals. Following large-scale migration to Greece in the 1990s, the number declined to 15,100 in 2002 and further to 5,500 in 2014 (see Loladze, 2021 for the dynamics of the Greek population in Georgia from 1800 to 2014).

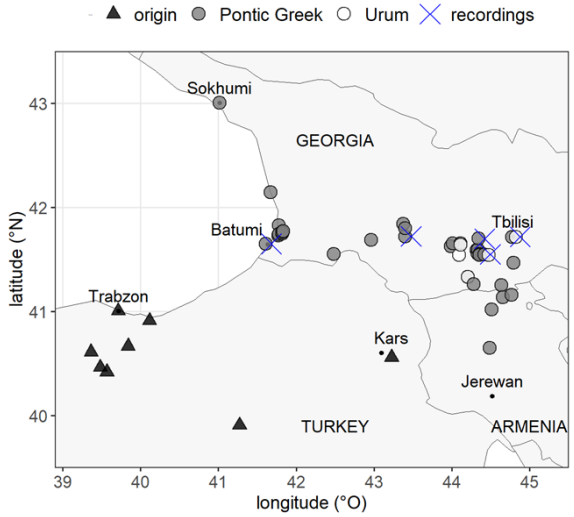


Figure 1: Greek populations in the Caucasus

Since Pontic Greek in the Caucasus originates from different locations of Anatolia, it is dialectally heterogeneous. However, the extent of koineization processes in the new settlements remains to be investigated. A common characteristic of these varieties is their development in a multilingual environment, shaped by sustained contact with Russian (since the nineteenth century), with Georgian (particularly in urban centers in recent decades), as well as with Standard Modern Greek, especially following renewed transnational contacts and labor migration since the 1990s. The corpus presented in this article has been designed to capture exactly these processes across different types of settlements.

3. Related Resources

Most available research and resources on Pontic Greek concern the varieties spoken in Anatolia, based either on data collected in Turkey or from emigrant speakers in Greece. The grammatical and historical dimensions of the dialect have been examined in comprehensive grammars, grammatical sketches and related studies (Papadopoulos, 1955; Oikonomidis, 1958; Mackridge, 1987; Tombaidis, 1988; Drettas, 1997; Revithiadou and Spyropoulos, 2012; Sitaridou, 2013, among many others).

Accessible digital databases containing primary Pontic data remain limited. A major resource is the AMiGre database (Pontus, Cappadocia, Aivali: in search of Asia Minor Greek),[1] which provides both raw and processed material from three dialects. The database includes an oral corpus of approx. 180h (with a multimodal sub-corpus of around 45h), a written corpus, and an electronic tri-dialectal dictionary.[2] Part of the corpus (15h per dialect) is annotated (Karanikolas et al., 2015).

Additional resources provide materials for special purposes. A dataset designed for NLP-based dialect identification includes Pontic (282,000 words) alongside other Greek dialects (Chatzikiriakidis et al., 2025). Pontic data are also available within the WOWA (Word Order in Western Asia) corpus, a spoken-language resource aimed at investigating areal effects in word order variation (Schreiber, 2021). Ethnographically collected Pontic data (without linguistic annotation) form part of the VLACH, Vanishing Languages and Cultural Heritage Collection. [3] Audio recordings representing various dialectal varieties, including Pontic (18 sound-files, 2 of which represent Pontic Greek in Georgia), are accessible through the Ulster University repository. [4] Large resources of written and spoken materials have also been used for the creation of the online atlas of Cappadocian Greek, DiCaDLand (Digitization of the Cappadocian Dialectal Landscape), which, however, does not include Pontic Greek (Melissaropoulou et al., 2022).[5]

Pontic Greek as spoken in the Caucasus remains comparatively under-resourced and less systematically investigated. Existing grammatical descriptions include Lazarev (1950), Eloeva (1997), Mikaberidze and Shakhpazidi (2000), and Berikashvili (2017) on Pontic Greek in Georgia as well as Hodgson (2010) on Pontic Greek in Armenia. However, comprehensive and digitally accessible resources for these varieties are still lacking.

4. Data Collection

4.1. Speakers

Participants were informed about the aims of the data collection and provided written informed consent for their data to be used in scientific research and deposited in online repositories. The speaker

[1] https://amigre.upatras.gr/

[2] http://amigredb.philology.upatras.gr

[3] https://www.oeaw.ac.at/vlach/collections/greek-varieties/pontic-greek

[4] https://www.ulster.ac.uk/research/topic/modern-languages-linguistics/quality/research-projects/investigating-variation-and-change

[5] http://cappadocian.upatras.gr/en/node/10

sample represents three (idealized) sociolinguistic settings:

- Stage A: HOMELAND: villages traditionally populated by Pontic Greek speakers: Manglisi, Tetritskaro, Orkhevi, Santa, Tsikhisjvari (see "recordings" in Figure 1).
- Stage B: INTERNAL MIGRATION: urban centers of Georgia, with Pontic Greek speakers living in strongly multilingual/multiethnic contexts: Tbilisi.
- Stage C: EXTERNAL MIGRATION: Pontic Greek speakers from Georgia who migrated to Greece: Thessaloniki (speakers originating from different places in Georgia).

Demographic information for each stage is summarized in Table 1. The smaller sample in stage B is due to the difficulty of identifying Pontic Greek speakers in urban centers such as Tbilisi.

stage	*n*	F	M	year of birth		
				min	max	mean
A	18	12	6	1928	1990	1958
B	8	4	4	1925	1989	1950
C	16	7	9	1939	1994	1975

Table 1: Speaker samples (total: *n*=42)

The metadata include information on the location and date of the recordings, as well as speaker-related variables: (a) gender and year of birth and (b) self-reported frequency of using Pontic Greek, Russian, Georgian, and Standard Modern Greek in the communication with family members and neighbors. Figure 2 presents aggregated results of self-assessed language use, which are informative for the vitality of the dialect in the sociolinguistic settings at issue. Speakers assess (on a 1-5 Likert scale) the frequency of use of Pontic Greek and the corresponding dominant language (Russian in stages A/B; Standard Modern Greek in stage C). Stage A speakers show balanced use of Pontic vs. Russian (dominant lg.), whereas the use of the dominant language (Russian or Greek) increases in stages B and C. In the latter stages, Pontic Greek is primarily used within the family and more frequently with the older generation (parents) than with children. Across stages, the use of Pontic Greek increases with social distance (more Pontic within the family) and with the age of the addressee (more Pontic when communicating with the older individuals, with the age effect being stronger within the family).

These self-estimated data confirm the rationale of sampling speakers across three stages, reflecting distinct sociolinguistic situations that influence language use.

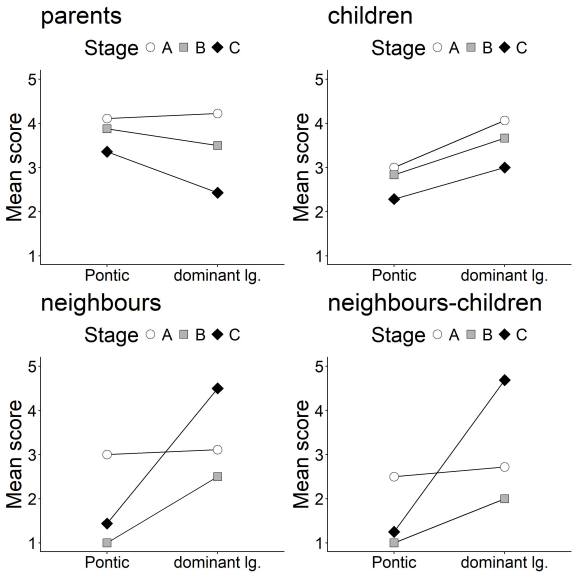


Figure 2: Aggregated self-estimation scores of the language use with different addressees

4.2. Texts

This corpus was designed specifically to study variation between speakers in the stages introduced in §4.1. To this end, all participants were instructed to produce the same eight narratives on the following topics: ANCESTORS, FAMILY, VILLAGE, CULTURE, PEOPLE, MARRIAGE, FEAST, and LANGUAGE. Four of the prompts are illustrated below; the remaining prompts are available in the online documentation of the resource:[6]

- ANCESTORS: How did your ancestors come to Georgia?
- COMPARISON: Please tell us how your people are different from the other people in the village/city?
- MARRIAGE: Please tell us how your people celebrate an engagement/marriage and what is the difference to the way other people in this village/city celebrate an engagement/marriage.
- CULTURE: 'Please tell me a fairy tale or a poem in your native language. If you do not know any fairy tale or poem, please tell me what you consider most important in the culture of your people.'

Consequently, this corpus offers a dataset for special purposes, which may not have the breadth to serve as the basis for a comprehensive grammatical description. For instance, the restriction to narrative texts already implies that certain speech acts,

[6] https://spw.uni-goettingen.de/projects/xtyp/PNT.html

such as directives, are only marginally represented. This corpus is suitable for investigating variation between individual speakers or speaker groups on the basis of linguistic phenomena that recur in narratives, such as the realization of specific phonemes, intonation patterns, frequent morphosyntactic categories such as case marking and definiteness, word order, etc. From an ethnographic perspective, the corpus provides information for the specific cultural practices described in the eight narratives.

Beyond the main text collection (Kotanidi et al., 2019), the online resource also includes additional texts collected during the project (Berikashvili, 2019; Berikashvili and Skopeteas, 2019), that are not discussed here in detail.

4.3. Data Collection Procedure

The interviews were conducted in non-laboratory environments. Speakers were instructed in Pontic Greek by a native speaker of the dialect. Before the questions on the individual narratives, speakers were received the general instruction:

'Please answer the following questions spontaneously. Speak naturally, as if you are speaking to a friend. It does not matter if you are unsure about details, just give a natural response.'

Recordings were made using an Olympus LS-11 recorder with built-in microphones, at a sampling rate of 44.1 kHz. Audio files were saved in .wav format.

5. Annotations

5.1. Data Processing

The recordings were transcribed and annotated by domain experts, combining native-speaker competence and linguistic expertise. Transcriptions were made by a speaker with competence in Pontic Greek, Russian and Georgian. Transcriptions were subsequently morphologically analyzed and glossed by a linguist with expertise in all related languages in Toolbox (SIL, 2015). Toolbox annotations were then exported to ELAN (Sloetjes and Wittenburg, 2008), where sentence boundaries were time-aligned with the audio files. The annotation is expert-based and was not designed for inter-annotator agreement, which is difficult to implement for endangered varieties with limited expert availability.

Project materials include soundfiles (.wav) and annotations in xml format (.eaf files, ELAN). All files follow a unified file-naming convention using the template: language-collection-task-additional-speaker. For example, the filename PNT-TXT-AN-00000-A03.wav is interpreted as shown in Table 2.

PNT	*Language identifier* for 'Pontic Greek' (ISO-639-3)
TXT	*Subcollection identifier* TXT= subcollection 'texts'
AN	*Task identifier* AN= 'ANCESTORS'
00000-	*Additional info* no information in this example
A03	*Speaker identifier* speaker 3, sample A (HOMELAND)

Table 2: File-naming template

5.2. Annotation Tiers

The corpus has a multi-layer annotation structure, including transcriptions, morphological segmentation, glossing, POS tagging, and translation. The annotation tiers (in Toolbox and ELAN files) are listed in Table 3. (The tiers marked with the suffix *_a* represent the original annotations, which are subsequently revised in the corresponding primary tiers.)

name	**parent**	**content**
tx	-	*text:* revised transcriptions
tx_a	ref	*text_associated:* original transcriptions
mb	tx	*morphemic boundaries:* transcription enriched with morphemic boundaries
ge	mb	*gloss-English:* revised morpheme-by-morpheme translation (English)
ge_a	mb	*gloss-English_ associated:* original word-by-word translation (English)
ps	mb	*part of speech:* part of speech classification
ft	ft	*free translation:* sentence-by-sentence translation (English)
nt	ref	*notes:* free comments

Table 3: Annotation tiers

5.3. Transcriptions

Text transcriptions ('tx' tier) were based on orthographic conventions that are generally close to a broad phonological transcription. Greek words were transcribed using a system based on the phonemic transcription with IPA-symbols, as proposed by Drettas (1997); see (1). The symbol *-ä*

represents the phonemic value [æ], appearing in cases of sinezesis (/ia/ > /æ/) as in (1b) or foreign words, as in (1c). A correspondence table (orthographic symbols to phonemic values) as well as further transcription conventions are listed in the corpus website.[7]

(1) a. *érðan* 'come:PFV.PST:3.PL'
b. *ospítä* 'house:N.PL.NGEN'
c. *tätä-m* 'father-CL.1.SG:GEN'

Embedded Russian elements are transcribed according to the BGN/PCGN romanization standard; see (2a). Georgian elements are rendered following the Georgian national system of romanization (Apridonidze and Chkhaidze, 2002); see (2b). When elements of Georgian/Russian origin are morphologically integrated into Pontic Greek (e.g., by inflection), they are transcribed with the orthographic conventions adopted for Pontic Greek (in order to avoid switching orthographies within a word); see (2c). Word stress is indicated in all words by an acute accent (´).

(2) a. *prinimát'* 'take:IPFV.INF'
b. *ch'ianch'véla* 'ant:SG.NOM'
c. *chixirtmás* 'Chikhirtma:F.PL.NGEN' (Georgian chicken soup)
Georgian romanization: *chikhirtma*

5.4. Morphological Annotations

Morphological annotations comprise morphological segmentation ('mb' tier) and morpheme-by-morpheme translations ('ge' tier). Morphological annotations were based on The Leipzig Glossing Rules [8] – with addition of abbreviations from Eurotyp (see König et al., 1993).

As a general practice for highly inflectional languages, inflectional forms of Pontic Greek were not segmented with morphemic boundaries (at the 'mb' tier), but morphemes are translated in the 'ge' tier according to category-specific templates, as illustrated in (3) for nominals (substantives, adjectives, pronouns, adjectival participles).

The template for verbs is presented in (4). By convention, unmarked categories (active, indicative, present, finite) are not represented in glossing.

(3) **Nominal template**
stem:{gender}.{number}.{case}
ángelos 'angel:M.SG.NOM'
eγó '1.SG:NOM'

Nominal categories are added to the verbal categories of participial forms.

(4) **Verbal template**
stem:{voice}.{mood}.{aspect}.{tense}.{finiteness}
érθan 'come:PFV.PST.3.PL'
efovéθen 'fear:MEDP.PFV.PST.3.SG'
éxo γraménon 'have:1.SG write:PTCP:N.SG.ACC'

Additional transcription conventions were introduced to account for phenomena characteristic of spoken communication, as illustrated in Table 4.

xxx	unidentified words (mb/ge layer)
HESIT	hesitation
((laughs))	laughing

Table 4: Miscellaneous

6. Online Resources

The resources of this corpus contain three subcollections:

- **Subcollection TXT:** 24 speakers, 338 sound files (duration: 6h 7m), 35,843 words (Kotanidi et al., 2019).
- **Subcollection VA1:** 4 speakers, 30 sound files (duration: 38m), 3,830 words (Berikashvili, 2019).
- **Subcollection VA2:** 2 speakers, 5 sound files (duration: 34m), 3,400 words (Berikashvili and Skopeteas, 2019).

Audio files and transcriptions are available for download from the TLA archive (*The Language Archive*[9]) at the Max Planck Institute for Psycholinguistics in Nijmegen, a repository dedicated to data collections from endangered languages.

The corpus website[10] contains an overview of the resources, including the instructions used in data collection and the conventions for orthographic and morphological transcriptions. Users can navigate the resource and the audio files through a server installation of ANNIS (see Fig. 3), which is linked to the corpus website (Krause and Zeldes, 2016).

[7] Abbreviations: 1= 1st person; 3 = 3rd person; ACC = accusative; CL = possessive clitic; F = feminine; GEN = genitive; INF = infinitive; IPFV = imperfective; M = masculine; MEDP = mediopassive; N = neuter; NGEN = non-genitive; NOM = nominative; PFV = perfective; PL = plural; PST = past; PTCP = participle; SG = singular.

[8] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

[9] https://archive.mpi.nl/tla/

[10] https://spw.uni-goettingen.de/projects/xtyp/PNT.html

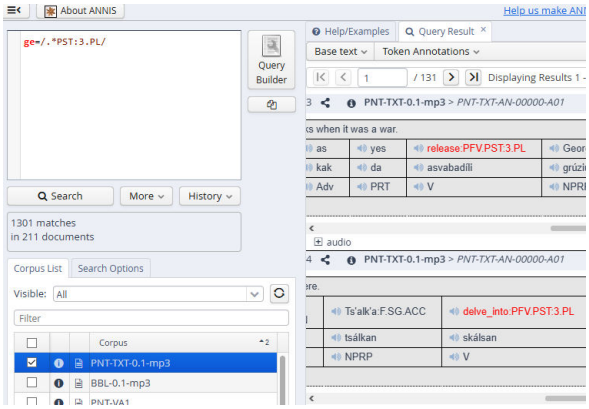


Figure 3: ANNIS interface for corpus queries

ANNIS provides the possibility to formulate queries over multi-layer annotations using its query language (AQL, ANNIS Query Language), which allows users to search for annotation attributes and relations using regular expressions. The illustrative query in Fig. 3 retrieves all tokens in the corpus that contain the string `PST:3.PL`, preceded by an arbitrary number (*) of arbitrary characters (.). Further example queries in AQL are provided on the corpus website.

7. Illustrating Variation in the Corpus Data

Pontic Greek people have always lived in multilingual settings, and the corpus reflects contact with several languages: (a) Turkish, including historical influence and later contact with Turkic varieties (e.g. Urum and Azerbaijani), especially in mixed regions and most strongly in texts from Dmanisi; (b) Standard Modern Greek, maintained through transnational family ties – particularly visible in Adjara; (c) Russian, with widespread bilingualism across regions; and (d) Georgian, which has become particularly influential following urban migration, especially to the capital.

The linguistic variation captured in the corpus is illustrated in the following examples. Figure 4 illustrates a text elicited with the CULTURE prompt (see §4.2). The excerpt, taken from a Pontic song, exhibits authentic Pontic lexical material and grammatical features, such as a subject in 'accusative' form: *xamelétan* 'mill:M.SG.ACC'.

The effects of language contact are also evident in lexical embeddings. Figure 5, for instance, demonstrates the embedding of a Russian noun in a Greek polydefinite. Currently, the language of origin (Russian, Turkish, Georgian) is only indicated in the Toolbox lexicon. In a latter version of our resource, this information can be exported to the ELAN files to be used in queries.

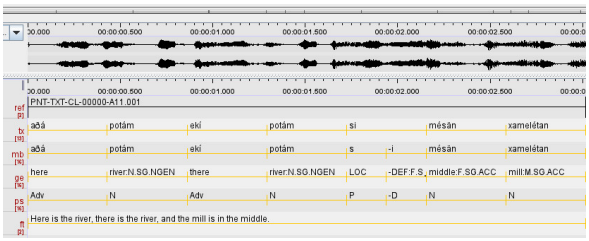


Figure 4: Pontic song (PNT-TXT-CL-00000-A11)

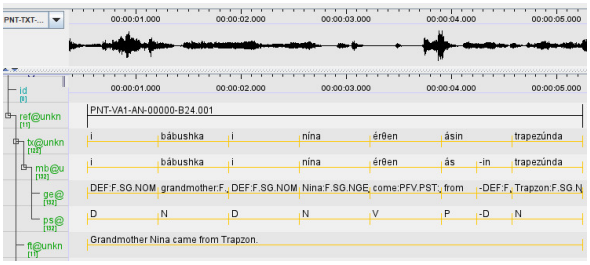


Figure 5: Russian embeddings in nominal structures (PNT-VA1-AN-00000-B24)

Figure 6 illustrates a typical strategy for integrating Russian verbs such as *prinimat'* 'take:IPFV.INF', namely through light-verb constructions (matrix verb: *epínen* 'make:IPFV.PST.3.SG'). The integration of verbs through light-verb constructions is cross-linguistically widespread (Wohlgemuth, 2009) and constitutes one of the main strategies for accommodating Russian loan verbs in Pontic Greek of Georgia (see Berikashvili, 2019 for a detailed discussion).

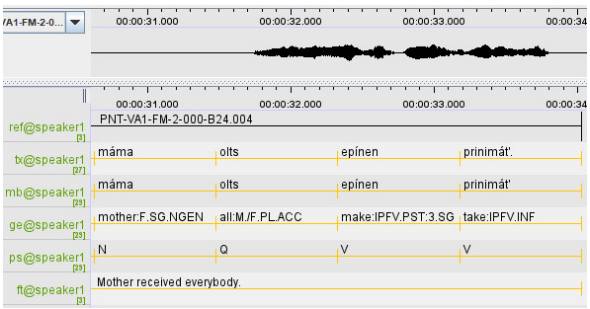


Figure 6: Russian embeddings in verbal structures (PNT-VA1-FM-2-000-B24)

The data in this corpus allows for quantitative assessments of effects of contact, such as the assignment of loanwords to inflectional classes or gender (see Berikashvili, 2022). Overall, the treatment of embeddings—both lexical and limited structural—places the variety represented in this corpus in the "more intense contact" category of the borrowing scale proposed by Thomason and Kaufman (1988).

8. Concluding Remarks

This article has presented an open-access, morphologically annotated multimedia corpus of Pontic Greek as spoken in the Caucasus, containing 373 recordings across three sociolinguistic stages that are representative for the settings in which the dialect is currently used. The resource fills a gap in the documentation of Pontic Greek outside Anatolia and contributes to our knowledge of an endangered dialect whose speaker base has drastically declined over recent decades. The corpus design – eliciting comparable narratives across speakers – facilitates systematic comparison between speakers and enables comparison between three important stages: homeland, internal migration, and external migration contexts.

The data highlights the complex multilingual ecology of the Caucasus, revealing patterns of contact with Russian, Georgian, Turkish, and Standard Modern Greek, and the resulting hybrid linguistic profiles. Beyond its relevance for Greek dialectology, the corpus provides valuable material for research on language contact, possible phenomena of koineization, and dialect attrition. We hope that this resource will serve both as a foundation for future empirical studies and as a contribution to the broader effort of safeguarding the linguistic heritage of Pontic communities in the Caucasus.

9. Acknowledgements

The corpus recordings, as well as the original transcriptions and translations of the main subcollection, were produced by Evgenia Kotanidi (Tbilisi). Svetlana Berikashvili revised the transcriptions and translations, added morphological analyses and glosses, and created two additional subcollections that form part of this corpus. Stefanie Böhm and Johanna Lorenz supervised the data collection and transcription process. The creation of this corpus was funded by the Volkswagen Foundation. We are grateful to two anonymous reviewers for their helpful comments and suggestions on an earlier version of this article. Responsibility for the present article rests with the authors.

10. Bibliographical References

Shukia Apridonidze and Levan Chkhaidze. 2002. From Georgian and into Georgian: Transliteration of Georgian alphabet. Linguistics Institute of Academy of Science of Georgia.

Svetlana Berikashvili. 2017. *Morphological Aspects of Pontic Greek Spoken in Georgia*. LINCOM, Munich.

Svetlana Berikashvili. 2019. Loan verb adaptation in Pontic Greek spoken in Georgia. In *Language contact in the Balkans and Asia Minor*, pages 262–279, Thessaloniki. Institute of Modern Greek Studies.

Svetlana Berikashvili. 2022. Contact-induced change in the domain of grammatical gender in Pontic Greek spoken in Georgia. *Languages*, 7:79.

Stergios Chatzikiriakidis, Chatrine Qwaider, Ilias Kolokousis, Christina Koula, Dimitris Papadakis, and Efthymia Sakellariou. 2025. Grdd: A dataset for Greek dialectal nlp. Version v5.

Georges Drettas. 1997. *Aspects pontiques*. Association de Recherches Pluridisciplinaires, Paris.

Fatima Eloeva. 1997. *Pontijskij dialekt (na materiale grecheskih bespis'mennyh govorov Gruzii i Krasnodarskogo kraja) [Pontic dialect (based on data of oral Greek varieities spoken in Georgia and Krasnodar Krai]*. St. Petersburg University, St. Petersburg.

Katherine Hodgson. 2010. Morphological evidence for the origins of a Pontic Greek community in armenia. *Studies in Greek Linguistics*, 30:237–248.

Concha Maria Höfler. 2020. *Boundaries and Belonging in the Greek Community in Georgia*. Nomos, Baden-Baden.

Nikitas N. Karanikolas, Eleni Galiotou, Dimitris Papazachariou, Konstantinos Athanasakos, George Koronakis, and Angela Ralli. 2015. Towards a computational processing of oral dialectal data. In *PCI '15: 19th Panhellenic Conference on Informatics*, pages 337–341.

Simon Kaukhchishvili. 1942. Berdznebis dasakhlebis ist'oria sakartveloshi [the history of Greek settlement in Georgia]. *Works of the A. Tsulukidze Kutaisi State Pedagogical Institute*, IV:219–239.

Simon Kaukhchishvili. 1946. Berdznebis dasakhlebis ist'oria sakartveloshi [the history of Greek settlement in Georgia]. *Works of the A. Tsulukidze Kutaisi State Pedagogical Institute*, VI:125–153.

Thomas Krause and Amir Zeldes. 2016. Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118—139.

Ekkehard König, Dik Bakker, Östen Dahl, Martin Haspelmath, Maria Koptjevskaja-Tamm, Christiam Lehman, and Anna Siwierska. 1993. *Eurotyp Guidelines*. European Science Foundation Programme in Language Typology.

Evstathios Lazarev. 1950. *Calkinskij dialekt grecheskogo jazyka (foneticheskij i morfologicheskij analiz) [Tsalka dialect of Greek (phonetic and morphological analysis)]*. Ph.D. thesis, University of Tbilisi.

Nika Loladze. 2021. *Migratory Movements of Georgia's Greek Community. The Impact of Current Socio-economic transformations*. Peter Lang, Berlin.

Peter Mackridge. 1987. Greek-speaking moslems of North-East Turkey: Prolegomena to a study of the Ophitic sub-dialect of Pontic. *Byzantine and Modern Greek Studies*, 11(1):115–137.

Dimitra Melissaropoulou, Stavros Bompolas, and Charalampos Tsimpouris. 2022. Digital cartography in the service of preservation of cultural linguistic heritage: Implementing the electronic dialectal atlas of Cappadocian Greek. *Scientific Culture*, 8(2):135–146.

Avtandil Mikaberidze and M. Shakhpazidi. 2000. Sakartveloshi mtskhovreb berdzenta dialekt'isatvis [about the dialect of Greeks living in Georgia]. In *Berdznebi sakartveloshi [Greeks in Georgia]*, pages 129–178. Logos, Tbilisi.

Dimosthenis Oikonomidis. 1958. *Grammatikī tīs ellinikīs dialektou tou Pontou [Grammar of the Greek Pontic dialect]*. Academy of Athens, Athens.

Anthimos A. Papadopoulos. 1955. *Istorikī grammatikī tīs Pontikīs dialektou [Historical grammar of the Pontic dialect]*. Committee for Pontic Studies, Athens.

Anthi Revithiadou and Vassilios Spyropoulos. 2012. *Ophitikī. Ptykhes tīs grammatikīs domīs mias pontiakīs dialektou [Ofitika. Aspects of grammatical structure of the Pontic dialect]*. Publishing house of brothers Kyriakidi, Thessaloniki.

Laurentia Schreiber. 2021. Pontic Greek (Romeyka). In G. Haig, D. Stilo, M. C. Doğan, and N. N. Schiborr, editors, *WOWA - Word Order in Western Asia. A spoken-language-based corpus for investigating areal effects in word order variation*. University of Bamberg, Bamberg.

Eleni Sideri. 2006. *The Greeks of the former Soviet Republic of Georgia. Memories and Practices of Diaspora*. Ph.D. thesis, University of London.

SIL. 2015. Field linguist's toolbox. [Computer software].

Ioanna Sitaridou. 2013. Greek-speaking enclaves in Pontus today: The documentation and revitalization of Romeyka. In M. C. Jones and S. Ogilive, editors, *Keeping languages alive, documentation, pedagogy, and revitalization*, pages 98–112. Cambridge University Press, Cambridge.

Han Sloetjes and Pitter Wittenburg. 2008. Annotation by category: Elan and iso dcr. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 816–820, Marakech, Morocco. European Language Resource Association (ELRA).

Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkley, Los Angeles, London.

Dimitris Tombaidis. 1988. *Ī Pontiakī dialektos. Dialektika charaktīristika, katataxī idiōmatōn, dialektika keimena [Pontic dialect. Dialectic characteristics, classification and texts]*. Committee for Pontic Studies, Athens.

Jan Wohlgemuth. 2009. *A Typology of Verbal Borrowings*. Mouton de Gruyter, Berlin.

Artemis Xanthopoulou-Kyriakou. 1997. Metoikesies ton ellinōn tou pontou pros tis xōres tou kavkasou (1829-arxes 20ou aiōna) [migration of Greeks from Pontos to Caucasian countries (1829—beginnings of the 20th century)]. In I. Khasiotis, editor, *Oi Ellīnes tīs Rōsias kai tīs Sovietikīs Enōsīs. Metoikesies kai ektopismoi. Organōsī kai ideologia. [Greeks from Russia and Soviet Union. Migration and expatriation. Organization and Ideology]*, pages 85–127. University Studio Press, Thessaloniki.

11. Language Resource References

Berikashvili, Svetlana. 2019. *Pontic Data Collection 2*. The Language Archive, Corpus resource, 3.0. PID https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3.

Berikashvili, Svetlana and Skopeteas, Stavros. 2019. *Pontic Data Collection 3*. The Language Archive, Corpus resource, 3.0. PID https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3.

Kotanidi, Evgenia and Berikashvili, Svetlana and Böhm, Stefanie and Lorenz, Johanna and Skopeteas, Stavros. 2019. *Pontic Data Collection*. The Language Archive, Corpus resource, 3.0. PID https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3.

Meaning Over Morphology: A Multi-Metric Benchmark of LLMs for Bangla Dialect Translation

Soumik Deb Niloy[1], **Subhey Sadi Rahman**[2], **Mahbub E Sobhani**[1,2], **Golam Rabiul Alam**[1], **Farig Sadeque**[1], **Md. Rezuwan Hassan**[1]

[1]BRAC University, Dhaka, Bangladesh, [2]United International University, Dhaka, Bangladesh

msobhani2410011@mscse.uiu.ac.bd | md.rezuwan.hassan@g.bracu.ac.bd

Abstract

Regional dialects of Bangla, such as Sylheti and Chittagonian, pose significant challenges for natural language processing due to their low-resource nature and substantial linguistic variation from standard Bangla. In this work, we present a systematic evaluation of eight open-source LLMs for translating fifteen distinct Bangla dialects into standard Bangla. To achieve this comprehensive coverage, we utilize a combination of established benchmarks and a novel dataset curated from an ongoing regional linguistic project. We assess model performance using a multi-metric framework that combines exact-match and error-rate evaluations such as, Averaged BLEU, WER, and CER with embedding-based semantic metrics including BERTScore, METEOR, and COMET. Additionally, we perform a detailed dialect-level linguistic analysis to identify the deep-seated structural, orthographic, and semantic barriers inherent to dialectal translation. Our study highlights the strengths and limitations of current open-source models, provides empirical insights for future dialect-aware fine-tuning, and contributes a reproducible benchmark for the research community.

Keywords: bangla dialects, low-resource language, translation, large language models

1. Introduction

Bangla exhibits immense geographic and linguistic diversity across a vast dialectal continuum, historically categorized into major groups like Rarhi, Varendri, Kamrupi, and Bangla (Chatterji, 1970). As illustrated in Figure 1[1], peripheral variants such as Sylheti, Chittagonian, and Rangpur differ significantly from standard Bangla (primarily Rarhi-based) in phonology, vocabulary, and morphology that they frequently border on mutual unintelligibility.

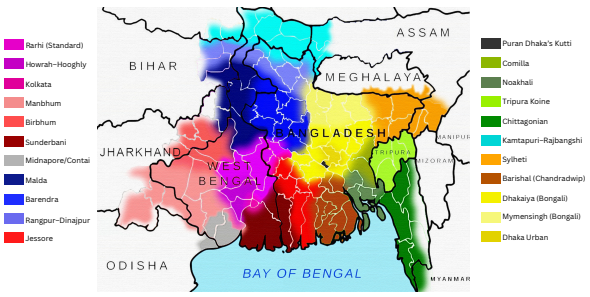


Figure 1: Regional dialects of the Bangla language.

This extreme variance creates a severe computational bottleneck. Table 5 in the appendix section provides examples that capture some of the regional dialectal diversity. IPA transcriptions are included solely for these instances to improve interpretability and are absent from the remaining data. Despite the rapid advancement of Large Language Models (LLMs), most foundation models are trained predominantly on standard Bangla, leaving these structurally different dialects largely underrepresented. Consequently, current Bangla NLP systems experience catastrophic performance drops often failing entirely at intent recognition, sentiment analysis, or semantic parsing when processing these peripheral variants. This failure carries a profound real world cost: millions of native dialect speakers are effectively excluded from modern digital inclusion initiatives, including voice assistants, e-governance platforms, and AI-driven healthcare, as their natural speech is systematically misinterpreted as erroneous. Translating these regional variants to standard Bangla is therefore an essential prerequisite for equitable downstream NLP performance.

However, this remains a fundamentally low-resource challenge characterized by a critical scarcity of aligned corpora. Furthermore, robust evaluation requires nuance. While the Averaged BiLingual Evaluation Understudy (BLEU) score serves as a widely adopted baseline for measuring n-gram translation precision, morphologically rich dialects demand evaluation methods that do not strictly penalize valid regional substitutions. The cost of relying exclusively on exact-match metrics

[1] Banerjee (2023), The Linguist Magazine

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 238–255
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

is substantial: it fundamentally misguides model development by artificially lowering the scores of models that successfully preserve semantic intent, thereby stalling progress in dialectal NLP. A comprehensive evaluation must therefore capture both exact-match surface normalization and deep semantic fidelity.

To address these limitations, we systematically evaluate open-source LLMs across fifteen distinct Bangla dialects. These specific dialects were strategically selected because they span the entire geographical and morphological continuum of Bangladesh capturing both mutually unintelligible linguistic islands (such as Chittagong and Sylhet) and central, transitional variants (such as Tangail and Mymensingh) thereby providing a holistic benchmark of true dialectal diversity. Leveraging a rigorously curated consolidation of existing public datasets, along with a private fold derived from an ongoing project referred to herein as the Regional-Standard Bangla Corpus (RSBC), we identify persistent structural translation barriers. Our evaluation employs a rigorous multi-metric approach. We combine surface-level lexical assessments such as Averaged BLEU (Papineni et al., 2002), Word Error Rate (WER), Character Error Rate (CER) with flexible and semantic-aware measures like Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Lavie and Agarwal, 2007), Crosslingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020), and BERTScore F1 (Zhang et al., 2020). Through detailed dialect-aware error analysis, this study provides a reproducible benchmark and critical insights, offering a roadmap for geographically inclusive, low-resource dialectal translation.

2. Related Work

The computational landscape for Bangla regional dialects is characterized by a significant disparity between data availability and task-specific utility. While dialects have been explored within socio-linguistic frameworks such as the *BIDWESH* (Fayaz et al., 2025) dataset introduced for multi-dialectal hate speech detection the development of parallel text corpora remains the most critical requirement for linguistic standardization and machine translation.

2.1. Evolution of Parallel Corpora and Translation Benchmarks

The primary obstacle to bridging dialects with standard Bangla is the lack of high-quality parallel data. Initial efforts to address this were often restricted to specific regions, such as *ChatgaiyyaAlap* (Chowdhury et al., 2025) for Chittagonian and *FeniVerse* (Mahi et al., 2025a) for the Feni dialect. Broader benchmarks like *Vashantor* (Faria et al., 2023) and the *BanglaDial* (Mahi et al., 2025b) corpus have since emerged to provide a multi-dialectal foundation. The *ONUBAD* (Sultana et al., 2025) dataset offers a comprehensive collection for automated conversion into standard Bangla, while recent works like *BhasaBodh* (Bhuiyan et al., 2025) provide parallel data for Chittagong and Sylhet, alongside addressing the real-world challenge of romanized digital communication. Despite these advancements, many sub-regional variations remain underrepresented, and the field continues to rely on a limited set of parallel pairs for model training.

2.2. From Traditional Neural Models to Large Language Models

Methodologically, dialect translation has shifted from rule-based mapping to Transformer-based architectures. Previous studies (Khandaker et al., 2025) have demonstrated that fine-tuned models like BanglaT5 can achieve a WER as low as 15.7% for standard-to-dialect translation.

Beyond text-based tasks, the complexity of Bangla dialects has also been investigated in the speech domain as well. For instance, Dipto et al. (2025) demonstrated that state-of-the-art ASR systems suffer severe performance drops on dialects like Chittagonian and Sylheti. Our benchmark mirrors this investigation in the text domain, evaluating whether LLMs exhibit the same vulnerability.

The current research frontier, however, centers on the deployment of LLMs. To adapt these foundation models, frameworks like *Sylheti-CAP* (Prama, 2025) have actively sought to improve zero-shot translation accuracy through sophisticated context-aware prompting. Systems like *BanglaDialecto* (Samin et al., 2024) have further attempted to unify speech recognition and translation into a single pipeline.

Building on this, recent work (Riad et al., 2025) has begun exploring parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) to develop dialect-aware conversational AI. For instance, Fine-tuned models such as Mistral, Gemma, and Claude for Question-Answering (QA) tasks across a subset of five regional variants. While these studies demonstrate the viability of fine-tuning LLMs for dialectal comprehension, they primarily rely on traditional n-gram metrics like BLEU and ROUGE for evaluation. As previously established, these exact-match metrics often fail to capture semantic fidelity in morphologically rich dialects. Furthermore, these efforts focus on chatbot-style QA generation rather than rigorous linguistic standardization. Consequently,

a comprehensive, multi-metric evaluation of open-source LLMs specifically targeting direct dialect-to-standard translation across a much wider array of geographically diverse Bangla dialects remains a critical gap in the literature.

3. Dataset

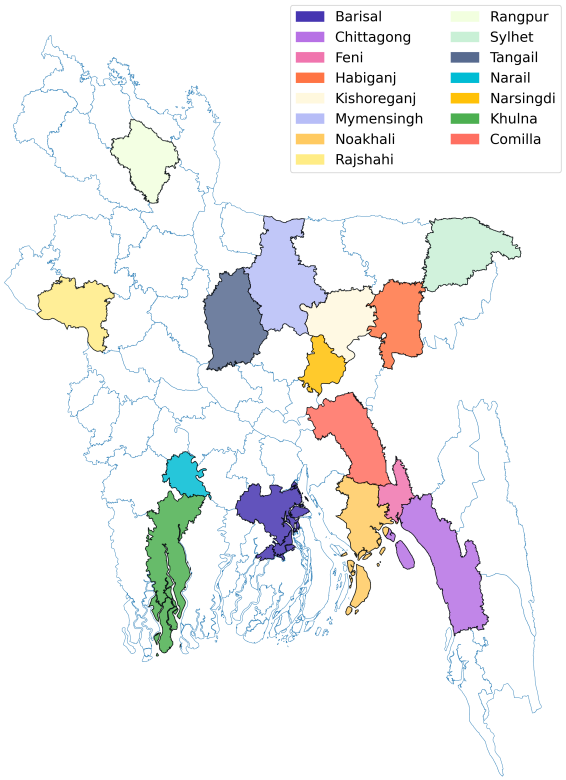


Figure 2: Map of evaluated dialects.

3.1. Task Definition

We formulate dialect-to-standard translation as a sequence-to-sequence generation task. A visual summary of the data conversion and preprocessing steps is presented in Figure 3. Given an input sequence of tokens in a regional Bangla dialect, the objective is to generate a semantically equivalent token sequence in Standard Bangla. The model must maximize the conditional probability of the standard translation given the dialectal input, ensuring deep semantic preservation while actively normalizing regional lexical, phonological, and morphological variations.

3.2. Proprietary Dialect Dataset

The RSBC corpus is composed of both aggregated open-source datasets and a private subset, which we designate as the Proprietary Dialect Dataset (PDD). This proprietary subset represents a curated checkpoint derived from a separate, ongoing research initiative dedicated to developing a large-scale Bangla parallel corpus. We integrated this private data into the current benchmark to augment our total sample size and expand our geographic coverage across a wider array of regions. While this specific subset remains temporarily proprietary pending the completion of its parent project, its inclusion ensures a more comprehensive, linguistically diverse, and statistically robust evaluation of the models. This PDD checkpoint currently encompasses data from 8 regions. To ensure high-quality parallel translations, we recruited 9 independent, native speakers for each respective regional dialect. The annotators were specifically trained for this task and spent approximately six months transcribing the data, while expert linguists continuously validated the accuracy of the outputs. All annotators were fairly compensated for their time.

3.3. Dataset Construction and Concatenation

To build a comprehensive multi-dialectal benchmark, we aggregated different publicly available sources such as *Vashantor* (Faria et al., 2023), *BIDWESH* (Fayaz et al., 2025), *Kothon* (Faisal et al., 2025), *BanglaDialecto* (Samin et al., 2024), *Bangla Regonal Text Corpus* (Ahmed et al., 2026), *Bangla Dialect Dataset* (Naeen et al., 2025), *FeniVerse* (Mahi et al., 2025a), *ChatgaiyyaAlap* (Chowdhury et al., 2025), *ONUBAD* (Sultana et al., 2025), in addition to the Proprietary Dialect Dataset (PDD) introduced in Section 3.2. In total, this benchmark encompasses 15 distinct regional Bangla dialects, including Chittagong, Noakhali, Barishal, Sylhet, Feni, Habiganj, Mymensingh, Khulna, Rangpur, Rajshahi, Tangail, Kishoreganj, Narail, Narsingdi, and Cumilla, as visualized in Figure 2. Each instance in the corpus provides a parallel sentence pair containing the original regional text, its Standard Bangla translation, and corresponding metadata tracking the region and dataset source. To ensure high data integrity across these disparate sources, all datasets were merged into a single unified corpus through a rigorous pipeline. This process included strict schema normalization, UTF-8 encoding standardization, a two-stage cross-dataset duplicate removal process utilizing both exact and fuzzy matching, and final manual and automated quality validation steps to remove malformed pairs.

The RSBC corpus contains 68,666 samples in total, with 17,004 derived from our proprietary dataset and 51,662 collected from external sources. Details regarding all these data sources

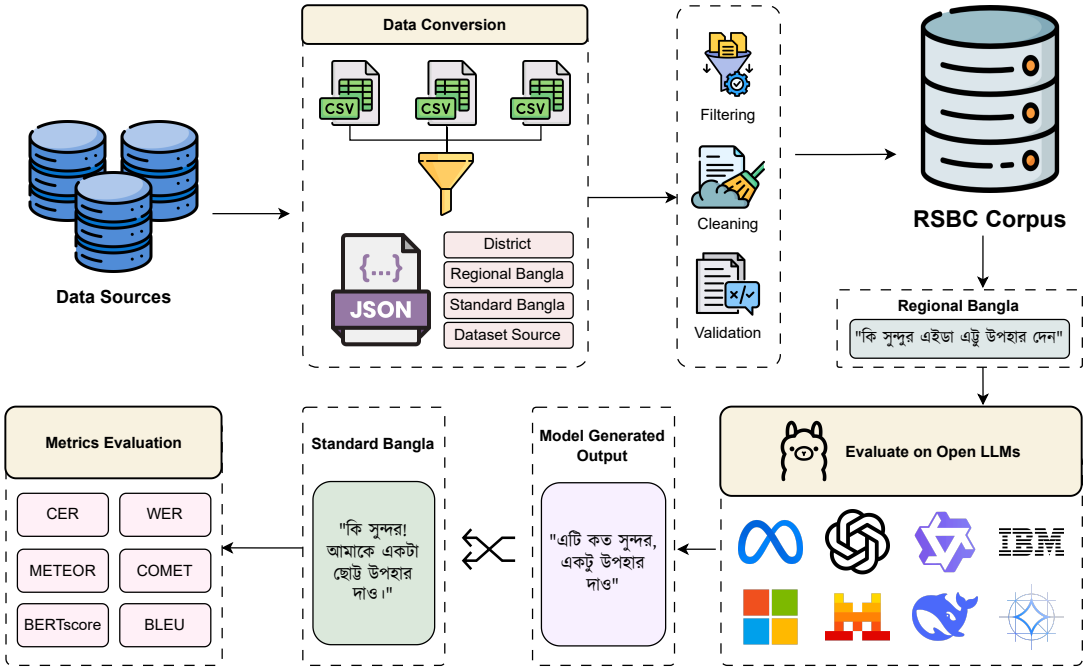


Figure 3: Overview of Regional–Standard Bangla Corpus (RSBC) creation and evaluation pipeline.

utilized in this work are presented in the Appendix (Section 12) at Table 3.

District	Samples	Total Words	Standard Unique	Dialect Unique	OOV (%)
Chittagong	19919	200524	19274	25225	23.59
Barishal	8065	80428	10386	11662	10.94
Noakhali	7983	89709	11492	13415	14.33
Rangpur	6108	40551	4677	7190	34.95
Habiganj	5357	50451	5564	8154	31.76
Sylhet	4904	36375	4305	6197	30.53
Feni	4035	42686	7865	8456	6.99
Kishoreganj	3623	21856	2844	4096	30.57
Mymensingh	2957	23467	3297	3927	16.04
Narsingdi	1973	14965	2416	3161	23.57
Narail	1143	6620	1467	1912	23.27
Khulna	909	4768	898	1266	29.07
Rajshahi	885	5624	1091	1621	32.70
Comilla	455	4145	768	1014	24.26
Tangail	350	2711	748	906	17.44
Total	**68,666**	**624,880**	**76,092**	**99,202**	**N/A**

Table 1: Dataset statistics and OOV% by district.

Table 1 summarizes the region-wise corpus statistics, including sample volume, vocabulary sizes, and the approximate Out-of-Vocabulary (OOV) percentage. Prior to analysis, texts were normalized and standard stopwords were removed.

Table 4 in the Appendix (Section 12) contains representative samples from our constructed RSBC corpus, illustrating regional dialect diversity and corresponding Standard Bangla translations. Figure 6 in the Appendix (Section 12) illustrates the distribution of samples across each regions.

As shown in Table 1, the aggregated corpus exhibits a severe class imbalance. High-resource dialects within our collection, such as Chittagong, contain nearly 20,000 samples, while the most low-resource variants, such as Tangail, contain only 350. A ratio of approximately 57:1. This disparity is an unavoidable reflection of the current scarcity of publicly available parallel corpora for peripheral Bangla dialects, necessitating reliance on the limited data currently available online and from our ongoing data collection efforts. Beyond sample distribution, two critical linguistic patterns emerge from this data.

First, for every evaluated region, the number of unique dialectal tokens significantly exceeds the unique tokens in the corresponding Standard Bangla translations (e.g., 25,225 dialect vs. 19,274 standard tokens in Chittagong). It is important to note that traditional Type-Token Ratio (TTR) is highly sensitive to corpus size, making cross-district comparisons of unique token counts (e.g., Chittagong vs. Tangail) susceptible to size-induced skew. However, because our corpus is strictly parallel, comparing the dialectal vocabulary directly against its Standard Bangla counterpart within the exact same set of sentence pairs perfectly controls for corpus size. This mathematically valid intra-dataset vocabulary inflation highlights the morphological richness and lack of standardized orthography in regional dialects, requiring models to learn complex, many-to-one mappings during translation.

Second, the OOV percentages reveal severe lexical bottlenecks. Regions such as Rangpur (34.95%), Rajshahi (32.70%), and Habiganj (31.76%) demonstrate that roughly one-third of their regional vocabulary is entirely disjoint from Standard Bangla.

This high prevalence of OOV tokens provides a direct, empirical explanation for why traditional

foundation models pre-trained almost exclusively on standard linguistic structures exhibit the elevated WER seen in Table 6 when attempting to normalize these peripheral variants.

To further quantify the structural and morphological divergence between the regional dialects and Standard Bangla, we analyzed the sentence length relationships across the parallel corpus. Figure 12 in the Appendix (Section 12) visualizes the character-level length distribution for all 15 evaluated dialects. In each subplot, a single data point represents a parallel sentence pair, plotted by its Standard Bangla length (x-axis) against its corresponding Regional Bangla length (y-axis). The solid diagonal line indicates absolute parity ($y = x$), where the dialect and standard sentences are identical in length.

While a strong positive correlation is naturally maintained across all districts, the visualizations reveal critical morphological variances. The dispersion of points away from the diagonal parity line illustrates that regional variations frequently utilize either highly compressed morphological constructs (falling below the line) or more verbose, descriptive phrasing (appearing above the line) to convey the same semantic meaning as the standard text.

Notably, highly divergent dialects such as Chittagong, Sylhet, and Noakhali exhibit broader scattering. Conversely, central and western dialects such as Tangail, Narail, Rajshahi, and Khulna demonstrate a much tighter clustering along the parity line. Specifically, the subplots for Rajshahi and Khulna reveal that their data points closely hug the diagonal, indicating that their sentence structures and morphological lengths strongly mirror Standard Bangla with minimal deviation. This variance mathematically demonstrates that translating peripheral Bangla dialects is not a symmetric, word-for-word substitution task. Instead, it inherently involves handling one-to-many and many-to-one token mappings. For foundation models (LLMs), this structural elasticity compounds the translation difficulty, as the model must bridge significant gaps in sentence length and syntactic structure to preserve the underlying intent, further reinforcing the necessity of semantic-aware evaluation metrics.

4. Dialect-Level Corpus Analysis

Beyond model evaluation, we conducted an exploratory linguistic analysis to quantify lexical and semantic variation across the fifteen regional dialects represented in our merged parallel corpus. Since this work is dialect-focused, understanding inter-dialect similarity is crucial for interpreting translation difficulty and model behavior.

4.1. Lexical Variation

To measure surface-level lexical overlap between dialects, we compute pairwise Jaccard similarity scores. For each region, we construct a vocabulary set consisting of unique word tokens extracted from the regional dialect side of the corpus.

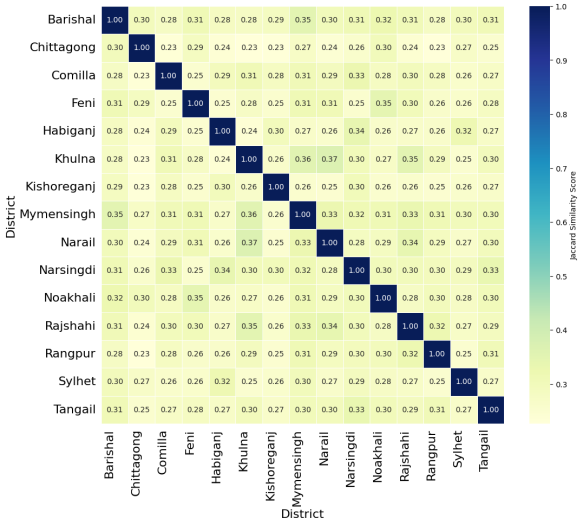


Figure 4: Jaccard Matrix of Bangla dialects.

The resulting Jaccard similarity matrix in Figure 4 reveals substantial lexical divergence across several geographically distant dialects, while neighboring regions tend to exhibit higher overlap. This supports the hypothesis that lexical variation contributes significantly to translation difficulty. While pairwise Jaccard similarity provides a useful heuristic for surface-level overlap, we note that the severe disparity in corpus sizes across districts (as outlined in Table 1) naturally suppresses the intersection ratio between high-resource and low-resource pairs. Consequently, these similarities represent corpus-level distributional overlap rather than absolute linguistic distance.

Specifically, the matrix demonstrates that Chittagong and Sylhet dialects exhibit the lowest similarity scores compared to Standard Bangla. These dialects function almost as linguistic islands, possessing distinct vocabularies and root words that severely limit surface-level overlap. In contrast, central and northern regions such as Tangail, Kishoreganj, and Rajshahi show markedly higher similarity indices, reflecting their closer morphological ties to the standard form.

4.2. Orthographic Divergence

To quantify the surface-level orthographic and morphological variation that models must overcome during normalization, we computed the average character-level Levenshtein distance between the sentences of each regional dialect and their corre-

sponding Standard Bangla translations. This metric measures the minimum number of character insertions, deletions, or substitutions required to transform one sentence into another, thereby capturing the spelling variations, affix differences, and phonological realizations inherent to each region.

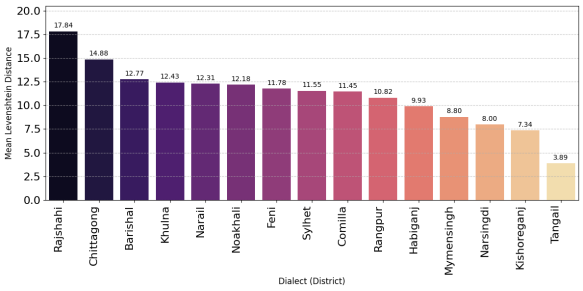


Figure 5: Average Character-Level Edit Distance to Standard Bangla.

The resulting bar chart in Figure 5 illustrates the mean edit distance to Standard Bangla across the evaluated regions. Higher values indicate a stronger orthographic divergence from the standard form. As shown, dialects from the southeastern and northeastern regions most notably Chittagong, Noakhali, Habiganj, and Sylhet exhibit the highest average edit distances. This reflects substantial phonological and morphological mutations that require heavy character-level alterations to standardize. In contrast, southwestern and central dialects such as Khulna, Narail, Cumilla, and Narsingdi demonstrate relatively lower edit distances, suggesting a much closer surface-form similarity.

Measuring the direct distance to the standard form provides a precise diagnostic of the normalization burden for each variant. This pronounced surface-level divergence offers a clear linguistic explanation for the elevated CER and WER observed in our zero-shot LLM evaluations, confirming that foundation models struggle heavily when confronted with severe morphological mutations, even if the underlying semantic intent is preserved.

4.3. Semantic Similarity Analysis

To evaluate meaning preservation and capture semantic closeness, we calculate the average cosine similarity between dialectal sentences and their Standard Bangla counterparts using multilingual Sentence-BERT embeddings (Reimers and Gurevych, 2019).

As reported in Table 2, similarity scores range from 0.632 in Chittagong to 0.897 in Tangail, revealing nuanced variations in semantic proximity. Regions such as Narail, Narsingdi, and Mymensingh also exhibit high alignment, showing scores of 0.822, 0.801, and 0.793, respectively. It suggests that they function as a linguistic "central hub" where core concepts and sentence structures remain highly preserved. In contrast, peripheral dialects like Chittagong, Sylhet, and Habiganj exhibit low semantic alignment, with scores falling to 0.632, 0.656, and 0.685, accordingly. This is particularly noteworthy because these regions are traditionally perceived as the most lexically distinct or even mutually unintelligible with Standard Bangla, a perception now validated by their positioning at the lower end of the semantic spectrum. These findings reinforce the critical research distinction between lexical divergence (surface-level word choice) and semantic divergence (the core meaning conveyed) when evaluating dialect-to-standard translation performance.

District	SSS	District	SSS
Tangail	0.896610	Rangpur	0.741638
Narail	0.821695	Kishoreganj	0.720103
Narsingdi	0.800928	Noakhali	0.716744
Mymensingh	0.793136	Rajshahi	0.688009
Khulna	0.782692	Comilla	0.686630
Feni	0.772380	Habiganj	0.685308
Barishal	0.743976	Sylhet	0.655511
Chittagong	0.632263		

Table 2: Average Semantic Similarity Scores (SSS) between regional dialects and standard Bangla.

Notably, these findings should be viewed through the lens of our dataset's sample distribution, which varies across regions. While the pronounced divergence in Chittagong and Sylhet highlights a significant linguistic gap, the high semantic proximity in Tangail suggests that these results serve as a high-signal indicator of how geographical proximity influences linguistic standardization. Ultimately, this observed semantic spread validates the hypothesis that translation difficulty is inherently dialect-dependent, necessitating more robust, dialect-aware modeling strategies to bridge these specific linguistic divides.

4.4. Lexical vs Semantic Correlation

To better understand how surface-level variation relates to deeper semantic alignment, we analyzed the correlation between lexical distance ($1 -$ Jaccard similarity) and sentence-level semantic similarity (Sentence-BERT). We observe a moderate negative correlation ($r = -0.47$) between lexical distance and semantic similarity across districts, indicating that increased lexical divergence generally leads to reduced semantic similarity. However, the relationship is not statistically significant ($p = 0.074$), suggesting that many dialectal variations preserve meaning despite substantial

lexical differences, see Appendix (Section 12.2) at Figure 7.

Rather than a weakness, this lack of strict statistical significance is a profound linguistic finding that perfectly supports our core hypothesis. As illustrated in Figure 7, while peripheral dialects (e.g: Chittagong or Sylhet) exhibit extreme vocabulary shifts, their semantic degradation is not perfectly linear. This proves that dialects are not simply "noisy" Standard Bangla, but robust, expressive linguistic systems capable of maintaining semantic integrity even when surface-level overlap is minimal. Consequently, this validates our argument that traditional exact-match metrics (which heavily penalize lexical divergence) are fundamentally inadequate for this task, and deep semantic-aware evaluations are absolutely necessary to judge true translation capabilities.

This corpus-level finding provides critical linguistic grounding for the observed LLM performance disparities: peripheral dialects do not merely suffer from surface-level spelling variations, but from deep-seated vocabulary shifts that fundamentally break the semantic mapping of standard foundation models. Consequently, this compounded divergence exponentially increases the translation burden, validating the severe spikes in WER observed in our evaluation. More importantly, it proves that standard zero-shot inference is fundamentally insufficient for bridging extreme dialectal gaps; resolving this semantic-lexical disconnect strictly necessitates the development of dedicated, dialect-aware fine-tuned LLMs capable of internalizing these robust regional linguistic structures.

5. Methodology

5.1. LLM Inference Framework

5.1.1. Problem Formulation

We are studying the task of translating regional Bangla text into standard Bangla across 15 distinct region dialects. Each problem instance consists of a regional Bangla input sequence, denoted as $R = \{r_1, r_2, \ldots, r_n\}$, where each r_i originates from a specific dialect within the region set $D = \{d_1, d_2, \ldots, d_{15}\}$. These inputs are paired with a fixed zero-shot instruction prompt, denoted as P_{zs}, which instructs the model to perform the dialect-to-standard conversion without relying on prior in-context examples. The objective is to generate the correct standard Bangla text, represented as $\hat{S}$, for each regional input r_i. This framework enables us to evaluate the inherent zero-shot translation capabilities of a LLM. Throughout the execution of our pipeline, the parameters for the model were kept entirely frozen. The entire procedure can mathematically be abbreviated as follows:

$$\hat{S}_i = LM([P_{zs}, r_i]) \quad (1)$$

Figure 3 illustrates a high-level overview of the proposed methodology.

5.1.2. Prompting Strategy

To ensure a robust and reproducible evaluation, we employ a structured zero-shot prompting framework (Kuo and Chen, 2023). Our methodology utilizes an instruction-based approach designed to enforce strict linguistic role conditioning and output constraints. Specifically, the prompt template incorporates structured tagging (e.g., `<translation>...</translation>`) to facilitate deterministic parsing and minimize both model hallucinations and extraneous conversational filler. To maintain experimental integrity and ensure a fair comparison across the 15 region-specific datasets, we apply an identical prompt configuration to all evaluated models. This standardized formatting ensures that performance variations are attributable to the models' inherent linguistic capabilities rather than prompt engineering biases. The comprehensive prompt template used throughout this study is detailed in Appendix 12.3.

5.2. Experimental Setup

5.2.1. Models

We evaluated a diverse suite of state-of-the-art LLMs to assess translation capabilities across various regional variants and dialects. Our selection encompasses both foundational and reasoning-optimized architectures, including `Llama3.2` (Grattafiori et al., 2024), `Phi4` (Abdin et al., 2024), `Qwen3` (Yang et al., 2025), `gpt-oss` (Agarwal et al., 2025), and `IBM Granite4` (Mishra et al., 2024). To further probe the limits of reasoning-heavy translation tasks, we included `DeepSeek-R1` (DeepSeek-AI, 2025), `Mistral` (Jiang et al., 2023), and `Gemma2` (Team et al., 2024). All models were deployed in a zero-shot prompting configuration to evaluate their inherent linguistic resilience and latent dialectal understanding without the influence of task-specific fine-tuning.

5.2.2. Evaluation Metrics

To rigorously quantify translation quality, we utilized a multi-metric framework designed to capture both lexical precision and semantic fidelity. For evaluating meaning preservation beyond surface-level overlap, we employed BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020), which

leverage contextual embeddings to ensure semantic alignment. We further utilized METEOR (Lavie and Agarwal, 2007) for its sensitivity to synonymy and stemming. To measure baseline n-gram translation precision, we report the averaged BLEU (Papineni et al., 2002) score, which captures structural mapping from unigrams up to 4-grams, while WER and CER were used to provide granular diagnostics of morphological and lexical normalization accuracy.

6. Result Analysis

6.1. Performance Analysis

The comprehensive evaluation detailed in Table 6 at Appendix (Section 12.4) reveals significant disparities in how current LLMs handle the structural and morphological complexities of regional Bangla. Specifically, the data highlights a severe gap between exact lexical normalization measured by BLEU, WER, and CER and the underlying semantic preservation captured by BERTScore and COMET.

`gpt-oss` and `Qwen3` emerge as the dominant performers across the macro-average. `gpt-oss` achieves the highest overall semantic alignment with a METEOR score of 0.3324, a COMET score of 0.8106 and a BERTScore of 0.8402, while also demonstrating the best exact-match BLEU average of 0.2224. `Qwen3` closely trails `gpt-oss` in semantic retention scoring a BERTScore of 0.8329, but secures the lowest overall CER of only 0.4333, indicating superior capability in handling fine-grained morphological standardizations. Both models showcase robust resilience when translating linguistically central dialects like Tangail, where `gpt-oss` reaches an impressive BERTScore of 0.8897 and a comparatively low WER of 0.6052.

However, it is crucial to interpret the macro-average reported in Table 6 at Appendix (Section 12.4) with caution due to the aforementioned dataset imbalance. The macro-average treats all fifteen districts equally, meaning the results from 350 samples of Tangail carry the same weight as those 19,919 samples from Chittagong. The exceptionally high performance figures for Tangail (e.g., `gpt-oss` achieving a BERTScore of 0.8897) likely reflect its closer linguistic proximity to Standard Bangla and its smaller, highly curated sample size, which is less representative of the true translation difficulty seen in heavily divergent dialects. Consequently, the unweighted macro-average may artificially inflate the perceived overall performance. While we report the macro-average to establish a baseline across all fifteen regions, future benchmarking on this dataset should consider weighted averages or isolated reporting of high-resource versus low-resource clusters to provide a fairer performance picture.

In contrast, `Mistral` and `Llama3.2` struggle profoundly with this dialectal normalization task. `Mistral` records the lowest macro-average across all positive metrics with a BLEU score of 0.0826, and COMET score of 0.5671. Additionally, it suffers from the highest global error rates and showed a WER of 1.2612, and CER of 1.0924. For heavily divergent regions such as Narsingdi and Kishoreganj, `Mistral`'s WER balloons to 1.6315 and 1.7807 respectively, suggesting massive token hallucination and structural breakdown.

To qualitatively investigate these extreme WER values, we analyzed Mistral's raw generation logs for these specific districts. We observed that instead of performing a direct mapping, Mistral frequently succumbed to severe output length mismatch driven by its failure to comprehend the peripheral dialect syntax. For example, given the Kishoreganj input "আন্নেরা দিয়া কারেন্ট আইছে?", the target Standard Bangla reference is "আপনাদের ঐদিকে কারেন্ট এসেছে?" (containing 4 tokens). However, Mistral completely failed the instruction and generated "আপনি এক্সট্রা হলো এক বিশ্ববর্তী ভাষা বিজ্ঞানী..." followed by a continuous, non-sensical loop of over 6,400 hallucinated tokens. This type of error manifesting as repetitive hallucination drastically inflates the insertion penalty component of the WER formula, mathematically pushing the score well above 1.0 and highlighting the brittleness of standard foundation models on low-resource regional variants.

`Llama3.2` fares only marginally better, indicating that without specialized fine-tuning, standard base models lack the necessary vocabulary and syntax mapping to bridge extreme regional variations. It is important to contextualize these results with respect to model size. The `gpt-oss (20B)` model possesses nearly seven times the parameter count of models like `Llama3.2 (3B)` and `Granite4 (3B)`. While `gpt-oss` demonstrates superior semantic retention, this is largely expected given its massive parameter advantage. The struggles of `Llama3.2` and `Granite4` highlight the limitations of highly compact models in low-resource dialectal mapping, rather than a fundamental architectural flaw.

Finally, the data reinforces our core linguistic hypothesis: exact-match metrics inherently penalize the necessary lexical substitutions required for dialect translation. Even the best performing model, `gpt-oss`, achieves a maximum average BLEU score of only 0.2224, yet its BERTScore average of 0.8402 proves that the fundamental meaning is

largely preserved.

Thus, while exact-match evaluations highlight the surface-level brittleness of LLMs, semantic-aware metrics are vital to accurately assessing true translational intent across the Bangla dialect continuum. The radar chart in Appendix (Section 12.2) at Figure 8 provides a holistic view of each model's multi-metric footprint. Because exact-match metrics inherently penalize dialectal normalizations, the "BLEU" axis remains universally contracted across all models. However, highly capable models like `gpt-oss` and `Qwen3` demonstrate a wide, expansive footprint toward the semantic axes, specifically BERTScore, COMET while successfully minimizing their stretch toward the error axes for instance, CER and WER, proving their superior capability in handling morphologically rich Bangla dialects.

Figure 9 and Figure 10 in Appendix (Section 12.2) separate the evaluation into positive semantic quality metrics including METEOR, COMET, BERTScore, BLEU and negative structural error metrics such as CER, WER. The quality distributions visually confirm `gpt-oss`'s dominance across all embedding-based metrics, closely followed by `Qwen3`. Conversely, the error distributions in Figure 10 highlight the extreme lexical brittleness of `Mistral` and `LLaMA 3.2`, which suffer from distinct spikes in both CER and WER, indicating frequent hallucinations and poor morphological mapping.

Figure 11 plots the models on a trade-off curve between semantic preservation (COMET) and surface-level structural degradation (WER). This scatter plot effectively clusters the models into three tiers of dialectal comprehension. `gpt-oss` and `Qwen3` occupy the optimal top-left quadrant (high semantic retention, low error), while `Mistral` is isolated in the bottom-right quadrant, representing a near-total failure to normalize the peripheral dialect structures effectively.

7. Future Work and Conclusion

While this framework successfully identifies critical linguistic barriers, current evaluations remain constrained by the severe geographical imbalance and representational disparity of available parallel data. Because we aggregated all currently available public data alongside our ongoing project, the resulting corpus suffers from heavy skew (e.g., Chittagong vs. Tangail). As an ongoing resource-building initiative, we are actively expanding our corpus to specifically target and over-sample these underrepresented sub-regional varieties to correct this imbalance. By using this expanded dataset, our objective is to fine-tune and publicly release a dedicated regional dialect-to-standard Bangla translation model, complete with comprehensive evaluation scripts and model cards. Ultimately, by providing both curated linguistic resources and an open-source baseline model, we aim to establish a sustainable infrastructure for dialect-aware NLP research and promote broader digital inclusion across all Bangla-speaking communities.

In this work, we presented a systematic evaluation of open-source LLMs across fifteen Bangla regional dialects, revealing significant semantic and structural challenges in dialect-to-standard translation. Through a multi-metric evaluation framework combining surface-level and semantic-aware, we provided a comprehensive analysis of current model capabilities and limitations in low-resource dialectal settings. By identifying these persistent translation gaps and proposing a unified benchmarking approach, this study contributes a foundational step toward more inclusive, linguistically grounded, and regionally representative language technologies.

8. Acknowledgement

The authors are deeply grateful to the annotators who have contributed to our past and present data collection initiatives. We extend a special thanks to those currently working on our private dataset; their contributions are vital to the ongoing expansion of regional Bangla dialect resources and have made this systematic evaluation possible. We also wish to thank the anonymous reviewers for their insightful and constructive feedback, which significantly strengthened the final version of this paper.

9. Ethical Statement

All datasets used in this study were obtained and handled in compliance with their respective licensing terms, with no personally identifiable information retained. The proprietary dataset was curated under appropriate oversight, with full acknowledgment of annotator contributions. Our work aims to support digital inclusion for underrepresented Bangla dialect communities, and we are committed to publicly releasing our benchmark and evaluation framework. The authors declare no conflicts of interest.

10. Bibliographical References

Marah I. Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R.

Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card.

Md. Tofael Ahmed Bhuiyan, Md. Abdur Rahman, and Abdul Kadar Muhammad Masum. 2025. BhasaBodh: Bridging Bangla dialects and Romanized forms through machine translation. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 113–118, Mumbai, India. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Tawsif Tashwar Dipto, Azmol Hossain, Rubayet Sabbir Faruque, Md. Rezuwan Hassan, Kanij Fatema, Tanmoy Shome, Ruwad Naswan, Md.Foriduzzaman Zihad, Mohaymen Ul Anam, Nazia Tasnim, Hasan Mahmud, Md Kamrul Hasan, Md. Mehedi Hasan Shawon, Farig Sadeque, and Tahsin Reasat. 2025. Are ASR foundation models generalized enough to capture features of regional dialects for low-resource languages? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 178–188, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and …. 2024. The llama 3 herd of models. *arXiv preprint*, 2407.21783.

Md. Rezuwan Hassan, Azmol Hossain, Kanij Fatema, Rubayet Sabbir Faruque, Tanmoy Shome, Ruwad Naswan, Trina Chakraborty, Md. Foriduzzaman Zihad, Tawsif Tashwar Dipto, Nazia Tasnim, Nazmuddoha Ansary, Md. Mehedi Hasan Shawon, Ahmed Imtiaz Humayun, Md. Golam Rabiul Alam, Farig Sadeque, and Asif Sushmit. 2025. Regspeech12: A regional corpus of bengali spontaneous speech across dialects.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Md. Arafat Alam Khandaker, Ziyan Shirin Raha, Bidyarthi Paul, and Tashreef Muhammad. 2025. Bridging dialects: Translating standard bangla to regional variants using neural models.

Hui-Chi Kuo and Yun-Nung Chen. 2023. Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 249–258, Toronto, Canada. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami Seelam, Brian Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David D. Cox, Ruchir Puri, and Rameswar Panda. 2024. Granite code models: A family of open foundation models for code intelligence. *arXiv preprint arXiv:2405.04324*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages

311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tabia Tanzin Prama. 2025. LLMs for low-resource dialect translation using context-aware prompting: A case study on Sylheti. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 292–308, Mumbai, India. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Md Jahid Alam Riad, Prosenjit Roy, Mahfuzur Rahman Shuvo, Nobanul Hasan, Stabak Das, Fateha Jannat Ayrin, Syeda Sadia Alam, Afsana Khan, Md Tanzim Reza, and Md Mizanur Rahman. 2025. Fine-tuning large language models for regional dialect comprehended question answering in bangla. In *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, and (and many others). 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, and Qwen3 technical report.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

11. Language Resource References

Ahmed, Md Tofael and Koli, Zannatul Mawa and Rahat, Azmain Mahtab and Akter, Taslima and Ayman, Umme. 2026. *BanglaRegionalTextCorpus: A curated dataset for four regional bangla dialects with standard Bangla and English translation*. PID https://doi.org/10.17632/92r62h4k5k.4.

Chatterji, Suniti Kumar. 1970. *The Origin and Development of the Bengali Language: Volume One*. Routledge, 1. PID https://doi.org/10.4324/9781003480945.

Chowdhury, Sinthia and Remal, Deawan Rakin Ahamed and Pasha, Syed Tangim and Islam, Ashraful and Noori, Sheak Rashed Haider. 2025. *ChatgaiyyaAlap: A dataset for conversion from Chittagonian dialect to standard Bangla*. Elsevier. PID https://doi.org/10.17632/wtms9xbkkw.1.

Faisal, Md. Atique and Sadaf, Farhan and Chowdhury, Dipta and Azrof, H.M. and Tanmay, Monojit Paul. 2025. *Kothon: A Large-Scale Dataset for Machine Translation of the Chittagonian and Sylheti Dialects into Standard Bangla*. Mendeley Data. PID https://doi.org/10.17632/2fv6vf9v2z.3. Version 3.

Faria, Fatema Tuj Johora and Moin, Mukaffi Bin and Wase, Ahmed Al and Ahmmed, Mehidi and Sani, Md. Rabius and Muhammad, Tashreef. 2023. *Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects to Bangla Language*. Mendeley Data. PID https://doi.org/10.17632/bj5jgk878b.2. ArXiv preprint arXiv:2311.11142.

Fayaz, Azizul Hakim and Uddin, MD. Shorif and Bhuiyan, Rayhan Uddin and Sultana, Zakia and Islam, Md. Samiul and Paul, Bidyarthi and Muhammad, Tashreef and Manzoor, Shahriar. 2025. *BIDWESH: A Bangla Regional Based Hate Speech Detection Dataset*. Mendeley Data. PID https://doi.org/10.17632/bpkrvf882k.1. ArXiv preprint arXiv:2507.16183.

Mahi, Mehraj Hossain and Khan, Anzir Rahman and Hoque, Zesanul and Mojumdar, Mayen Uddin. 2025a. *FeniVerse: A parallel corpus of Feni dialect, standard Bengali, and English*. Elsevier. PID https://doi.org/10.17632/25r923frfj.1.

Mahi, Mehraj Hossain and Khan, Anzir Rahman and Mojumdar, Mayen Uddin. 2025b. *BanglaDial: A merged and imbalanced text dataset for Bengali regional dialect analysis*. Elsevier. PID https://doi.org/10.17632/sm63ryv5dt.1.

Naeen, Md. Julkar and Das, Sourav Kumar and Supta, Supta Das Dip and Alahi, Tabib E

and Faisal, Md Faisal Tajwar and Rahat, Abdullah Al and Ayshe, Samurtha Jahan and Mahjabeen, Ummay and Tahmid, Md Tanvir and Mojumdar, Mayen Uddin. 2025. *Bangla Dialect Dataset: Exploring Linguistic Diversity Across Regions*. Mendeley Data. PID https://doi.org/10.17632/sm63ryv5dt.

Samin, Md. Nazmus Sadat and Ahad, Jawad Ibn and Medha, Tanjila Ahmed and Rahman, Fuad and Amin, Mohammad Ruhul and Mohammed, Nabeel and Rahman, Shafin. 2024. *BanglaDialecto: An End-to-End AI-Powered Regional Speech Standardization*. IEEE. PID https://doi.org/10.1109/BigData62323.2024.10826131.

Sultana, Nusrat and Yasmin, Rumana and Mallik, Bijon and Uddin, Mohammad Shorif. 2025. *ONUBAD: A comprehensive dataset for automated conversion of Bangla regional dialects into standard Bengali dialect*. Elsevier. PID https://doi.org/10.17632/6ft99kf89b.2.

12. Appendix

12.1. Dataset Sample Distribution

Dataset Source	Total Count	Covered Regions
Proprietary Dialect Dataset	17,004	Chittagong, Comilla, Habiganj, Kishoreganj, Narsingdi, Noakhali, Rangpur, Tangail
Vashantor	9892	Barishal, Chittagong, Mymensingh, Noakhali, Sylhet
BIDWESH	8985	Barishal, Chittagong, Noakhali
Kothon	7989	Chittagong
BanglaDialecto	4477	Noakhali
Bangla Regonal Text Corpus	4462	Barishal, Khulna, Narail, Rangpur
Bangla Dialect Dataset	4038	Barishal, Chittagong, Mymensingh, Rajshahi, Rangpur, Sylhet
Feniverse	4035	Feni
ChatgaiyyaAlap	4007	Chittagong
Onubad	2565	Barishal, Chittagong, Sylhet
Sylheti Dialects into Standard	1188	Sylhet
SylhetiCAP	24	Sylhet
Total	**68,666**	**12**

Table 3: Dataset-wise total sample count and covered districts (sorted in descending order).

Regions	Regional_bangla	Standard_bangla
Chittagong	আই হইদদে, হারে দইজজেদে?	আমি বলি যে কাকে ধরেছে?
Feni	হেতেরগো তুন অ্যাঁন্ডা ভালা খেইলছিলাম।	তাদের থেকে আমরা ভালো খেলেছিলাম।
Kishoreganj	ও আইচ্চা, তোরার বাইত গেলে কি খাইতারবাম?	ও আচ্ছা, তোদের বাড়িতে গেলে কি খেতে পারবো?
Sylhet	অউ ফুড়ি রীতিমতো মোর ভাই আর আম্মারে খষ্ট দের।	এই মেয়ে রীতিমতো আমার ভাই আর মাকে কষ্ট দিচ্ছে।
Mymensing	আমার কাল্কা একটা ভালা দাতের ডাক্তর দেহানির কথা।	আমার আগামীকাল একজন দাঁতের ডাক্তারের সাথে দেখা করার কথা।
Narsingdi	কিন্তু অয় যহন ঢুকছে দলে তহন থেইকা কিন্তু খুব ভালো খেলছে।	কিন্তু ও যখন ঢুকেছে দলে তখন থেকে খুব ভালো খেলছে।।
Barishal	হ্যায় যেমনে মোরে আশা করছে মুই নিজেরে হেমনেই সামনে আনছি।	সে যেভাবে আমাকে আশা করছে আমি নিজেকে সেভাবে উপস্থাপন করছি।
Khulna	কুথায় কুথায় ইংরেজি কবে আমাইগে সামনে হোডা আমাইগে সহ্য হবে না।	কখনও কখনও আমাদের ভাষার সাথে ইংরেজি মিশিয়ে ফেলা সহ্য করা হয় না।।
Rangpur	আজকা মুই তো হনু যে, আল্লাহর রাস্তায় যাইম, চল্লিশ দিন থাকিম, এই করিম।	আজকে আমি তো আল্লাহ্র রাস্তায় যাব, চল্লিশ দিন থাকব, এসবই করব।
Habiganj	তারফর ফাইয়াজের ফর ওইলো গিয়া আরেকটা আছে, ইটার নামটা বুইল্লা গেছিগা।	তারপর ফাইয়াজের পর হলো গিয়ে আরেকটা আছে এটার নাম ভুলে গিয়েছি।
Tangail	মধুপুরে রাবার বাগানের পরেই আছে মধুপুরের বিখ্যাত আনারসের হাট জলছত্র হাট।	মধুপুরে রাবার বাগানের পরেই আছে মধুপুরের বিখ্যাত আনারসের হাট, জলছত্র হাট।
Cumilla	এই কুত্তার বাচ্চা ধরতাম গিয়া আফনের অন্তত দুই থেকে তিন বার কুত্তার কামড় খাইছি।	এই কুকুরের বাচ্চা ধরতে গিয়ে আপনার অন্তত দুই থেকে তিন বার কুকুরের কামড় খেয়েছি।
Rajshashi	যখন সে হেবি একটা বেড নিউজ পেইয়ে গেলো তখন থেকেই সে খুব দুঃখ হোয়ে গেলো।	যখন সে খারাপ সংবাদ পেল, তখন তার দুঃখ হলো।
Noakhali	দুনিয়ার বুকে যেন দেওয়ান বাগির বিচার হয় যেন টাডা পরে আললা তোয়ার গজব নাজিল করে দাও।	দুনিয়ার বুকে যেন দেওয়ান বাগির বিচার হয় যেন টাডা পরে আললা তুমার গজব নাজিল করে দাও।
Narail	তুইতো ইবলিসকে আছিস রে ভাই! তোর খোঁজ করছিল ইবলিস। তোর নম্বরটা আই ইবলিসকে দি দিচি।	তুইতো ইবলিসকে আছিস রে ভাই! তোর খোঁজ করছিল ইবলিস। তোর নম্বরটা আমি ইবলিসকে দিয়েদিলাম

Table 4: Region-wise sample data from the RSBC corpus (sorted by Regional Bangla length)

Subset	Bengali text	IPA Transcription
Standard Bengali	সামনে ইদ, ইদের কেনাকাটা করতে হবে ভাই!	ʃɐmne ɪd̪, ɪd̪er kenɐkatɐ kor̪te hɔbe bʱɐ͡ɪ!
Rangpur	সামনোত ইদ, ইদের কেনাকাটি কইরবার নাগবে, ভাই!	ʃɐmnoʈo ɪd̪, ɪd̪er kenɐkatɪ ko͡ɪr̪bar nɐgbe, bʱɐ͡ɪ!
Kishoreganj	সামনে ঈদ, ঈদের কিনাকাডা করন লাগবো ভাই!	ʃɐmne ɪd̪, ɪd̪er kɪnɐkɐdɐ kɔron lɐgbo bʱɐ͡ɪ!
Narail	ছামনে ঈদ, ঈদির কিনাকাটা ওত্তি অবে ভাই!	cʰɐmne ɪd̪, ɪd̪ɪr kɪnɐkɐtɐ oʈʈ̪ɪ ɔbe bʱɐ͡ɪ!
Chittagong	সামনো দি ঈদ, ঈদর কিনাকাটা গরন ফরিবু বাই!	ʃɐmno d̪ɪ ɪd̪, ɪd̪or kɪnɐkɐtɐ gɔron pʰorɪbu bɐ͡i̯!
Narsingdi	সামনে ইদ, ইদের কিনাকাডা করুন লাগবো ভাই!	ʃɐmne ɪd̪, ɪd̪er kɪnɐkɐdɐ korun lɐgbo bʱɐ͡i̯!
Tangail	সামনে ইদ, ইদের কিনাকাটা করন নাগবো ভাই!	ʃɐmne ɪd̪, ɪdeɾ kenɐkɐtɐ kɔron nɐgbe bʱɐ͡ɪ
Habiganj	সামনে ইদ, ইদের লাইগ্গা কেনাখাডা খরোন লাগবো বাই!	ʃɐmne ɪd̪, ɪdeɾ lɐ͡ɪggɐ kenɐkʰadɐ kʰɔɾon lɐgbo bɐ͡i̯!
Sylhet	সামনে ইদ। ইদোর কেনাকাটা করা লাগবো, বাই।	ʃɐmne ɪd̪, ɪdoɾ kenɐkɐtɐ koɾɐ lɐgbo bɐ͡i̯.
Barishal	হোমনে ইদ, ইদের কেনাকাডা করতে অইবে বাই।	homne ɪd̪, ɪdeɾ kenɐkɐdɐ korte o͡ibe bɐ͡i̯.
Sandwip	সামনে ইদ, ইদের কিনাকাডা কইরতো হইব বাই।	ʃɐmne ɪd̪, ɪdeɾ kɪnɐkɐdɐ ko͡ɪr̪to ho͡ɪbo bɐ͡i̯.
Cumilla	সামনেদো ইদ, ইদের কিনাকাডা করতইবো বাই।	ʃɐmneɖo ɪd̪, ɪd̪er kɪnɐkɐdɐ korʈo͡ɪbe bɐ͡i̯.
Noakhali	সামনে ঈদ আইয়্যের, ঈদের বাজার হদাই করন লাইগবো ভাই।	ʃɐmne ɪd̪ a͡ɪer, ɪdeɾ baɟɐɾ hɔdoɪ kɔron lɐ͡ɪgbo bʱɐ͡i̯.

Table 5: Dialect diversification from different regions taken from (Hassan et al., 2025)

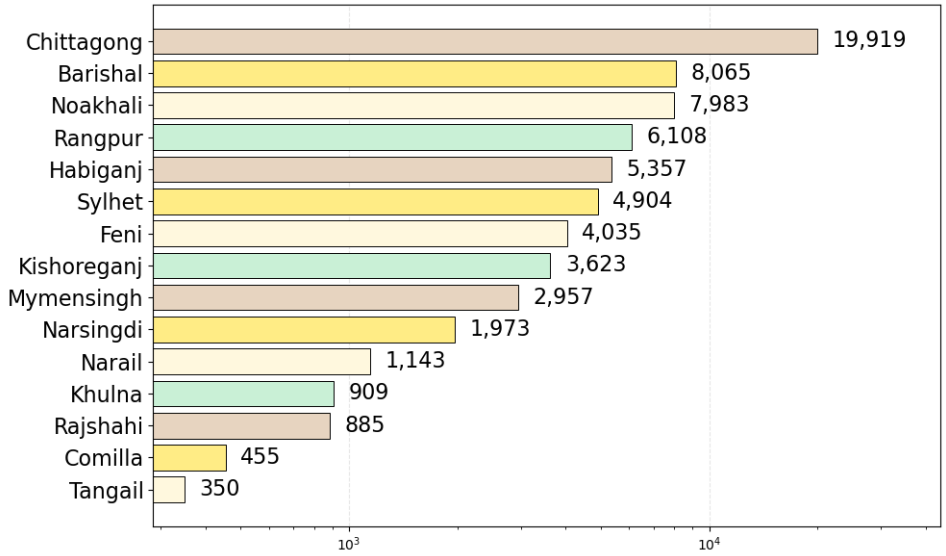


Figure 6: region-wise sample distribution.

12.2. Dialect-level Corpus Analysis

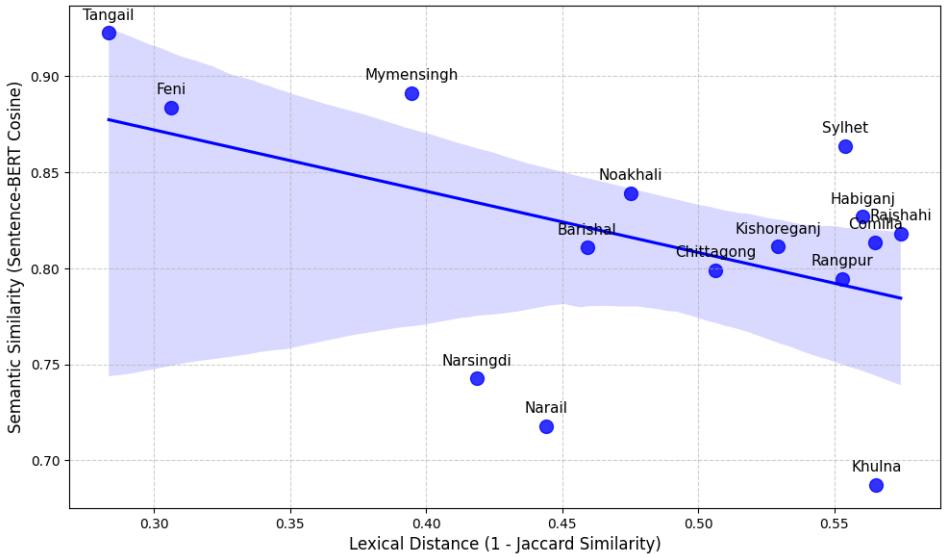


Figure 7: Correlation between Lexical Distance and Semantic Similarity across all the evaluated dialects

12.3. Prompt

The following prompt was utilized as the system instruction:

```
You are an expert linguist specializing in Bangladeshi regional
   dialects and Standard Bangla.

TASK: Translate the following {district} regional dialect text
into Standard Bangla (প্রমিত বাংলা).

CRITICAL INSTRUCTIONS:
- Provide ONLY the Standard Bangla translation.
- Do NOT provide any explanation, reasoning, or analysis.
- Do NOT include the original text in your response.
- Do NOT add any prefixes like "Translation:" or "Standard Bangla:".
- Output ONLY the translated text in Standard Bangla.
- Wrap your final translation in <translation>...</translation>
   tags.

REGIONAL DIALECT INPUT ({district}):
{regional_text}

STANDARD BANGLA TRANSLATION:
```

Here, `{district}` dynamically specifies the dialect source (e.g., Chittagong, Sylhet, Noakhali), and `{regional_text}` represents the dialectal input sentence.

12.4. Performance Analysis

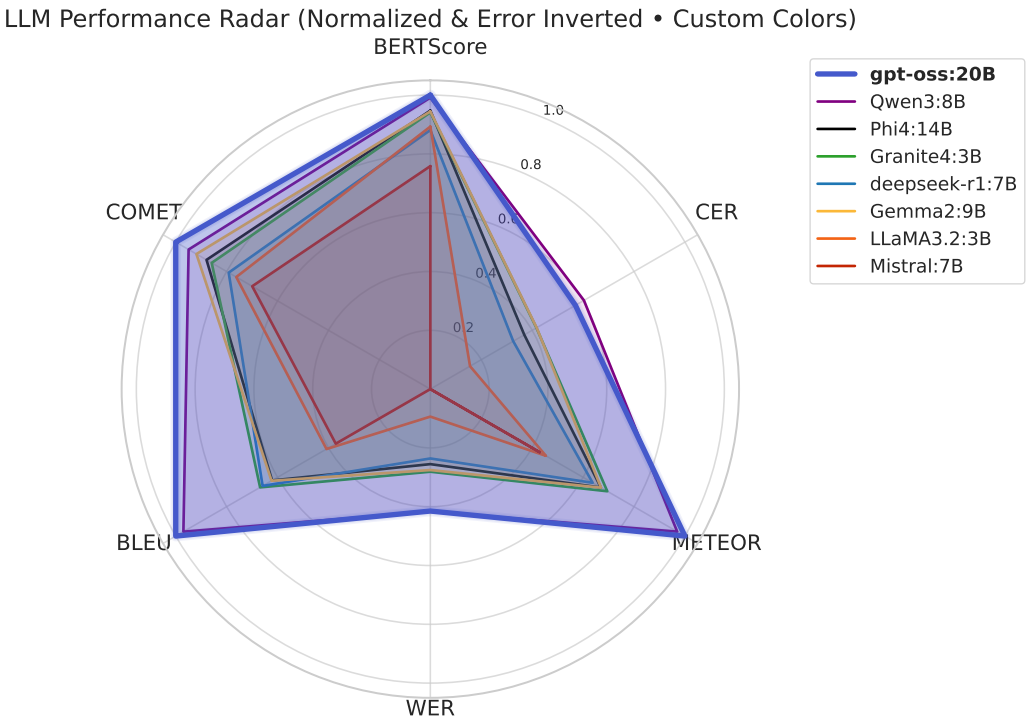


Figure 8: Multi-metric radar chart depicting the holistic performance footprint of each LLM.

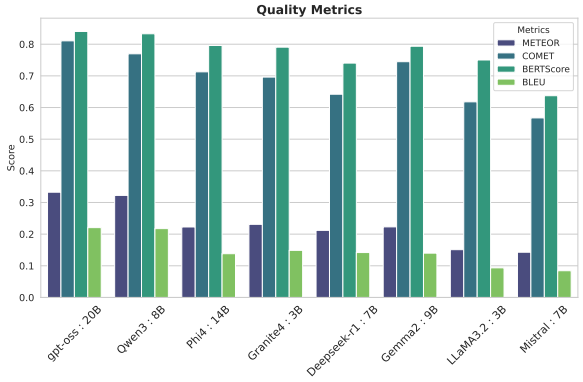


Figure 9: Macro-average distribution of quality metrics across evaluated LLMs.

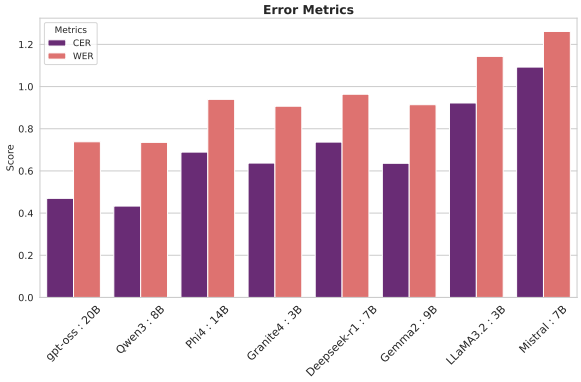


Figure 10: Macro-average distribution of error metrics across evaluated LLMs.

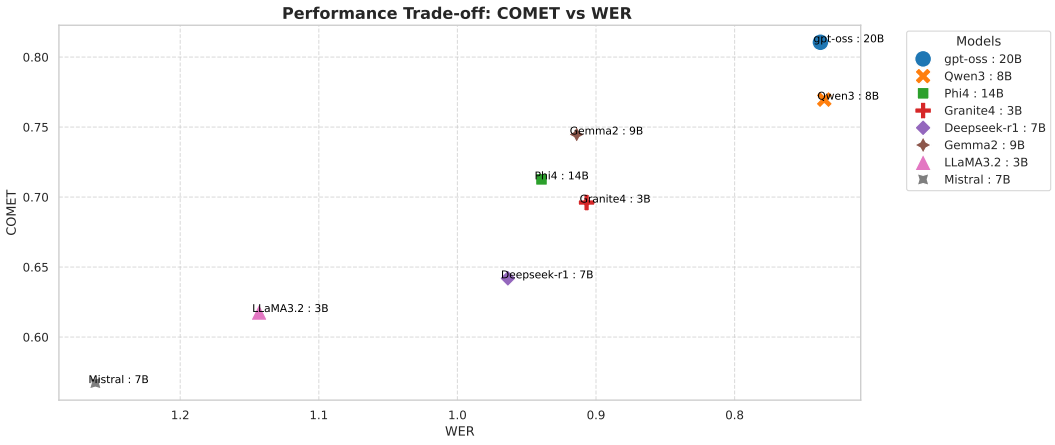


Figure 11: Performance Trade-off: Semantic alignment (COMET) versus morphological error (WER).

Models	Barishal						Chittagong					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.3893	0.8310	0.8515	0.2742	0.3957	0.6497	0.2712	0.7356	0.8245	0.1650	0.5537	0.8664
Qwen3:8B	0.3534	0.7782	0.8339	0.2471	0.4014	0.6786	0.2663	0.6796	0.8150	0.1797	0.4546	0.7728
Phi4:14B	0.2139	0.7067	0.7820	0.1366	0.6821	0.9313	0.1799	0.6422	0.7756	0.1111	0.7262	0.9764
Granite4:3B	0.2074	0.6283	0.7259	0.1416	0.6884	0.9123	0.1877	0.6145	0.7751	0.1206	0.6765	0.9519
deepseek-r1:7B	0.2399	0.6964	0.7933	0.1591	0.6125	0.8743	0.1586	0.5580	0.7257	0.1030	0.7682	1.0197
Gemma2:9B	0.2365	0.7431	0.7799	0.1556	0.6110	0.8787	0.1740	0.6697	0.7667	0.1099	0.7064	0.9763
LLaMA3.2:3B	0.1504	0.6046	0.7454	0.0961	0.9278	1.1340	0.1189	0.5532	0.7349	0.0724	0.9866	1.1912
Mistral:7B	0.1365	0.5514	0.6052	0.0848	1.0174	1.1803	0.1072	0.5095	0.5698	0.0640	1.1667	1.3058

Models	Cumilla						Feni					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2929	0.7511	0.8346	0.1787	0.4745	0.7699	0.3623	0.7983	0.8364	0.2409	0.3918	0.6807
Qwen3:8B	0.3043	0.7140	0.8415	0.1898	0.3932	0.7347	0.4261	0.7847	0.8504	0.3004	0.3483	0.6025
Phi4:14B	0.1933	0.6489	0.7863	0.1082	0.6855	0.9692	0.3156	0.7534	0.8182	0.2069	0.5041	0.7754
Granite4:3B	0.2237	0.6581	0.8009	0.1340	0.5786	0.8818	0.3055	0.7135	0.8094	0.2042	0.5164	0.7914
deepseek-r1:7B	0.2297	0.6310	0.7600	0.1451	0.7282	0.9617	0.3254	0.6573	0.7577	0.2296	0.5765	0.7716
Gemma2:9B	0.1884	0.6726	0.7775	0.1057	0.6767	0.9844	0.2625	0.7657	0.8010	0.1693	0.5546	0.8403
LLaMA3.2:3B	0.1308	0.5857	0.7485	0.0747	0.8689	1.1138	0.2490	0.6483	0.7772	0.1602	0.6993	0.9544
Mistral:7B	0.1410	0.5291	0.5858	0.0790	1.1389	1.2609	0.2022	0.5773	0.6593	0.1247	0.7252	0.9667

Models	Habiganj						Khulna					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2835	0.7482	0.8341	0.1624	0.5108	0.8163	0.3581	0.8921	0.8491	0.2423	0.4344	0.7071
Qwen3:8B	0.3163	0.7055	0.8381	0.1885	0.4074	0.7638	0.2718	0.8441	0.8083	0.1886	0.4981	0.8102
Phi4:14B	0.2020	0.6603	0.7892	0.1060	0.7139	0.9857	0.2578	0.7919	0.8274	0.1775	0.6122	0.8589
Granite4:3B	0.2317	0.6533	0.7966	0.1277	0.6288	0.9337	0.2013	0.7574	0.7877	0.1333	0.6601	0.9299
deepseek-r1:7B	0.2313	0.6083	0.7589	0.1359	0.7192	0.9853	0.1417	0.6832	0.7287	0.0986	0.7491	1.0051
Gemma2:9B	0.2038	0.6942	0.7896	0.1103	0.6734	0.9722	0.2763	0.8286	0.8290	0.1894	0.5412	0.8258
LLaMA3.2:3B	0.1529	0.5871	0.7460	0.0779	0.9211	1.1735	0.1160	0.6502	0.7346	0.0813	0.9189	1.1537
Mistral:7B	0.1553	0.5470	0.6174	0.0790	1.0835	1.2776	0.1255	0.6015	0.6820	0.0791	0.9470	1.1804

Models	Kishoreganj						Mymensingh					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2538	0.7509	0.8290	0.1353	0.5944	0.8570	0.4779	0.8827	0.8748	0.3535	0.3298	0.5498
Qwen3:8B	0.2746	0.7328	0.8392	0.1510	0.4746	0.8547	0.4425	0.8184	0.8630	0.3177	0.3520	0.5801
Phi4:14B	0.1661	0.6625	0.7894	0.0704	0.8746	1.1357	0.3011	0.7678	0.8070	0.2002	0.5395	0.7761
Granite4:3B	0.2039	0.6735	0.7977	0.1002	0.7104	1.0182	0.3146	0.7436	0.8210	0.2161	0.5150	0.7533
deepseek-r1:7B	0.1916	0.6267	0.7419	0.1039	0.9402	1.1339	0.2662	0.6399	0.7445	0.1853	0.6430	0.8386
Gemma2:9B	0.1757	0.6970	0.7883	0.0799	0.7253	1.0402	0.2935	0.8022	0.8099	0.1988	0.5241	0.7780
LLaMA3.2:3B	0.1141	0.5927	0.7295	0.0482	1.1160	1.3354	0.2350	0.6585	0.7861	0.1560	0.7177	0.9278
Mistral:7B	0.1328	0.5636	0.5971	0.0542	1.6930	1.7807	0.1871	0.5769	0.6810	0.1188	0.8730	1.0497

Models	Narail						Narsingdi					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.3822	0.9075	0.8583	0.2665	0.4030	0.6580	0.2456	0.7967	0.8270	0.1592	0.4869	0.8326
Qwen3:8B	0.3055	0.8684	0.8235	0.2141	0.4542	0.7494	0.2575	0.7746	0.8058	0.1869	0.4145	0.7858
Phi4:14B	0.2941	0.8198	0.8410	0.1994	0.5760	0.8179	0.1538	0.6770	0.7816	0.0960	0.7752	1.0661
Granite4:3B	0.2453	0.7960	0.7982	0.1632	0.6368	0.9042	0.1679	0.6935	0.7703	0.1138	0.6857	0.9929
deepseek-r1:7B	0.1878	0.7238	0.7407	0.1311	0.7198	0.9787	0.1856	0.6521	0.7181	0.1341	0.7861	1.0208
Gemma2:9B	0.3077	0.8530	0.8393	0.2069	0.5309	0.7920	0.1585	0.7113	0.7800	0.1007	0.6698	0.9841
LLaMA3.2:3B	0.1472	0.6932	0.7558	0.1016	0.8309	1.0796	0.1111	0.6118	0.7277	0.0741	1.1189	1.3586
Mistral:7B	0.1213	0.6035	0.6349	0.0785	0.9595	1.1718	0.1167	0.5870	0.6432	0.0759	1.3655	1.6315

Models	Noakhali						Rangpur					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2728	0.7464	0.8060	0.1740	0.4841	0.7921	0.3375	0.8279	0.8441	0.2141	0.4610	0.7521
Qwen3:8B	0.3240	0.7299	0.8274	0.2214	0.3661	0.6837	0.3033	0.7741	0.8314	0.1932	0.4699	0.8031
Phi4:14B	0.1836	0.6576	0.7664	0.1135	0.6575	0.9271	0.2185	0.7258	0.8033	0.1256	0.7375	0.9971
Granite4:3B	0.2140	0.6488	0.7799	0.1383	0.5916	0.8730	0.2302	0.7159	0.8006	0.1387	0.6454	0.9459
deepseek-r1:7B	0.2083	0.6013	0.6999	0.1392	0.6520	0.8842	0.1819	0.6327	0.7315	0.1124	0.7879	1.0360
Gemma2:9B	0.1853	0.6887	0.7599	0.1165	0.6463	0.9324	0.2243	0.7567	0.8023	0.1324	0.6567	0.9502
LLaMA3.2:3B	0.1513	0.5780	0.7453	0.0949	0.8810	1.0899	0.1198	0.6109	0.7382	0.0664	1.0389	1.2561
Mistral:7B	0.1396	0.5323	0.6062	0.0854	0.8956	1.0965	0.1359	0.5812	0.6472	0.0727	1.1609	1.3468

Models	Rajshahi						Sylhet					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2894	0.8385	0.8117	0.1938	0.5615	0.7809	0.3107	0.7970	0.8324	0.2180	0.5306	0.7533
Qwen3:8B	0.2702	0.7810	0.8063	0.1843	0.5832	0.8208	0.2785	0.7310	0.8240	0.1908	0.4738	0.7593
Phi4:14B	0.1961	0.7406	0.7728	0.1273	0.7174	0.9381	0.1914	0.6862	0.7748	0.1232	0.7865	1.0028
Granite4:3B	0.1878	0.7134	0.7752	0.1254	0.7201	0.9499	0.1957	0.6641	0.7861	0.1266	0.7233	0.9630
deepseek-r1:7B	0.1400	0.6180	0.7080	0.0926	0.9205	1.0803	0.1419	0.5772	0.7052	0.0953	0.8471	1.0445
Gemma2:9B	0.2034	0.7892	0.7762	0.1337	0.6670	0.9136	0.2126	0.7337	0.7840	0.1418	0.6464	0.8958
LLaMA3.2:3B	0.1379	0.6345	0.7505	0.0894	0.8842	1.0913	0.1311	0.5912	0.7513	0.0848	1.0082	1.2087
Mistral:7B	0.1147	0.5787	0.6716	0.0737	1.0092	1.1852	0.1208	0.5516	0.6648	0.0753	1.0815	1.2430

Models	Tangail						Average					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.4594	0.8546	0.8897	0.3282	0.4359	0.6052	**0.3324**	**0.8106**	**0.8402**	**0.2224**	0.4699	0.7381
Qwen3:8B	0.4409	0.8282	0.8852	0.3093	0.4079	0.6300	0.3223	0.7696	0.8329	0.2157	**0.4333**	**0.7353**
Phi4:14B	0.2714	0.7489	0.8234	0.1700	0.7497	0.9322	0.2226	0.7126	0.7959	0.1382	0.6892	0.9393
Granite4:3B	0.3482	0.7673	0.8349	0.2473	0.5808	0.8005	0.2310	0.6961	0.7906	0.1487	0.6372	0.9068
deepseek-r1:7B	0.3488	0.7237	0.7907	0.2574	0.5995	0.8189	0.2119	0.6420	0.7403	0.1465	0.7367	0.9636
Gemma2:9B	0.2407	0.7648	0.8195	0.1496	0.7121	0.9449	0.2229	0.7447	0.7935	0.1387	0.6361	0.9139
LLaMA3.2:3B	0.1977	0.6687	0.7806	0.1267	0.9129	1.0782	0.1509	0.6179	0.7501	0.0907	0.9221	1.1431
Mistral:7B	0.2059	0.6162	0.6930	0.1223	1.2685	1.2414	0.1428	0.5671	0.6372	0.0826	1.0924	1.2612

Table 6: District-wise and macro-average performance of LLMs for dialect-to-standard Bangla translation across fifteen districts. The best average metrics are in bold.

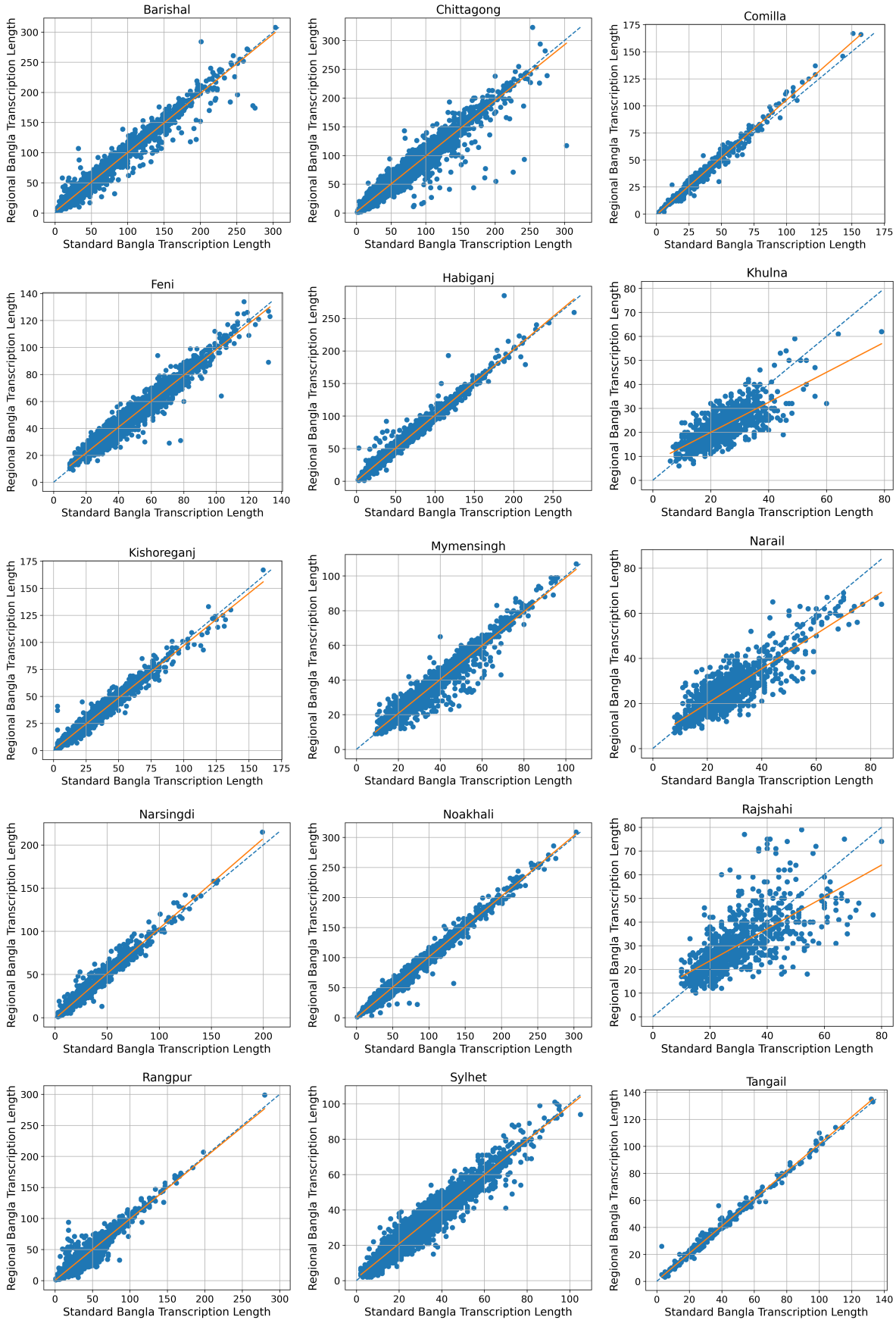


Figure 12: Sentence length comparison between Regional and Standard Bangla across dialects.

Sociolinguistic aspects of crowdsourcing for a vocal corpus of Alsatian

Pascale Erhart[1], **Lucile Hamm**[1], **Carole Werner**[1]
Malek Yaich[2], **Sam Bigeard**[2], **Slim Ouni**[2]
[1] LiLPa, Université de Strasbourg, Strasbourg, France
[2] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
pascale.erhart@unistra.fr lucilehamm@unistra.fr wernerc@unistra.fr
malek.yaich@inria.fr sam.bigeard@inria.fr slim.ouni@loria.fr

Abstract

Alsatian is a regional low-resource language spoken in a majority-language context. In order to create a voice dataset suited for training automatic speech recognition and speech-to-text models, we launched a crowdsourcing campaign on the platform Mozilla Common Voice. We describe sociolinguistic issues we ran into, such as participants' perception of their own language and its role in the AI landscape, which are vital to address to raise the participation in the crowdsourcing effort. We found that the participants are often confused about NLP and AI tools, and have a strong interested in preserving their language.

Keywords: Alsatian, Regional languages of France, German dialects, Crowdsourcing, Sociolinguistics

1. Introduction

Alsatian is a regional low-resource language spoken in a majority-language context: it is a German dialect spoken in France, where no language other than French benefit from any statutory recognition. Alsatian is one of the languages for which the COLaF project (Corpus et Outils pour les Langues de France, Corpus and Tools for the Languages of France) aims to contribute to the development of free corpora and tools. For Alsatian more specifically, the project aims to develop automatic speech recognition (ASR) and speech to text (TTS) tools, which requires speech and text training data in Alsatian. In order to overcome the usual issues associated with low-resource languages, such as limited and poor-quality data, the creation of a new corpus was undertaken via a crowsdsourcing campaign. For crowdsourcing to be successful, it is vital to take into account the public's perception of the topic. Thus, in this paper, we report on the sociolinguistic aspects and issues we encountered during the launch of the campaign, and what we did to address them.

The main contributions of this article are as follows:

- We give a detailed description of Alsatian and its sociolinguistic context to highlight the specific challenges encountered in developing NLP tools for this language (Section 2).
- We provide a state of the art in the field of Automatic Speech Recognition in Alsatian to show why the creation of a specific corpus is necessary (Section 3)
- We describe and discuss the different challenges we met during the implementation of the crowdsourcing campaign (Section 4).

2. Language context

The name "Alsatian" (als. *Elsässisch* or *Elsassisch*) is frequently used today to refer to the Alemannic and Franconian dialects spoken in Alsace since the 5th century AD. In this section, we describe the main characteristics of this low-resource language as well as its sociolinguistic context.

2.1. Dialect variation

The Alsatian dialect area is structured by phonetic elements which have their origins in the linguistic history of the German dialect area. The Alsatian dialects are mainly part of Upper German and partly part of Central German, which, from a geolinguistic point of view, make up the High German group.

Although it is characterized by variations in pronunciation as well as morphological or lexical variations, the Alsatian dialect area displays features that are shared by nearly all variants and allows its distinction from other German dialect areas. Among these features listed by Huck (2022), we can mention the regular palatalization of Middle High German (MHG) û (*Müs* [mys] 'mouse'), the lowering of MHG ë into [a], except in the northwest, the far north and along the Rhine up to the south of Strasbourg (*Laawe* [la:və] 'life'), the fact that /b, d, g/ consonants are realized as [b̥, d̥, g̊], that is to say as weak voiceless occlusives, or the de-rounding of certain rounded palatal vowels as in *scheen* [ʃe:n] ('beautiful') or *mied* [mi.ət] ('tired'). This

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 256–264
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

phenomenon affects nearly all the Alsatian area, whereas in Swiss German dialects, for example, labialization is maintained.

2.2. Effects of language contact

Alsace is a border region that was part of the German Empire for many centuries before becoming French at the end of the 17th century. During the last two centuries, it has been a theater for the conflicts between France and Germany and changed its state affiliation four times in less than 100 years (1871, 1919, 1940, 1945). Since 1945, Alsace is part of the French Republic, which officially recognizes French as its sole legitimate and official language. Since 2008, the French Constitution considers regional languages as heritage languages, but they are not subject to any specific policy. [1] This general context explains, at least partially, why the number of Alsatian speakers has declined since 1945. In the most recent survey conducted at the regional level Collectivité européenne d'Alsace (2022), 46 % of respondents stated that they 'speak Alsatian fairly well or very well', but three-quarters of these speakers were over 60 years old. The survey also reveals that understanding Alsatian is more important than speaking it, leading to a loss of the language.

Furthermore, it is important to note that nowadays all Alsatian speakers are also speakers of French, and often of other languages as well. This means that the use of Alsatian can be replaced by another language anytime and that speakers can use the multilingual resources of their repertoire in any spoken or written utterance in Alsatian. Language contact effects are thus numerous even in sentences that could be categorized as « in Alsatian ». Effects of language contact, especially with French, affect the morphosyntactic as well as the lexical level of the Alsatian dialects (Koehler, 2024) and make them diverge from other German dialects like Swiss German (although the High Alemannic dialects spoken in Alsace remain very close to the Swiss German dialects on the phonetic level) Koehler (2024). This explains why we cannot only draw on research on any other German dialect (like Swiss German or Bavarian) and need specific datasets for the finetuning of ASR or TTS models for Alsatian.

2.3. Graphic variation

At the local level, in Alsace, it was primarily a political vision, based on the idea that only a standard language could be taught in schools, that led the then head of the regional education system, Rector Pierre Deyon, to define German as the "regional language" of Alsace. The Académie de Strasbourg's current definition of "regional language" is as follows: « The regional language of Alsace is understood to be German in its standard form and in its dialectal variants (Alemannic and Franconian)» [2]. It is noteworthy that the term "Alsatian" is absent from this definition. Although both variants (dialect and standard German) are officially recognized as the "langue régionale d'Alsace", standard German is almost exclusively taught in public schools, with the exception of experimental trilingual programmes (French-Alsatian-German) introduced in 2023, which involve a very small number of pupils Hamm (2024).

When the opportunity arose in the 1980's to introduce a "bilingual" education in French and regional languages, « it was virtually impossible in Alsace to reach agreement on whether a bilingual education policy meant teaching both "French" and "Alsatian", or whether it should mean teaching "French" and "German", as "Alsatian" was not identified as a "language" » Le Page and Tabouret-Keller (1985). This view of Alsatian as a non-language still prevails and explains why it is still virtually absent from education today. Because of this lack of institutional recognition, Alsatian not being subject to any language policy, no intervention has been made regarding the corpus of Alsatian, so that the written corpus of this language is limited and characterized by graphical variation and the lack of a standard spelling. Its spelling is not normalized and reflects geolinguistic variation. This lack of institutionalisation also explains why Alsatian doesn't appear in language classifications like Glottolog Hammarström et al. (2025), where it generally falls under the umbrella of Swiss Alemannic, even if all Alsatian dialects are not part of the Alemannic area. [3]

Until the beginning of the 21st century, the use of written Alsatian was limited to regional literature, and more specifically to theater and poetry, so that "common/lay" speakers are not used to read or write in Alsatian. This changed with the arrival of digital communication and new writing habits : many people use Alsatian in their digital communication, especially on social media, which allows to observe how speakers deal with language variation in their written use of Alsatian : the absence of standard spelling systems leads to interpersonal variation,

[1] Constitutional Law n° 2008-724 of July 23, 2008 on the modernization of the institutions of the Fifth Republic, article 75 1 : « les langues régionales appartiennent au patrimoine de la France »

[2] Agreement on the regional multilingual policy, period 2015-2030, signed by the representatives of the State, the Région Alsace, the Département du Bas-Rhin and the Département du Haut-Rhin

[3] ISO 639-3 documentation

in which each writer chooses their own spelling convention Erhart (2026) Erhart (2025). The ORTHAL method Zeidler and Crévenat-Werner (2008) that was created to help normalizing the writing Alsatian is not well known by the general public, since it is not supported by any other institution than the local association who developed it. It can nevertheless be useful to normalize the written data gathered and to prepare them for their phonetisation.

3. Automatic Speech Recognition for Alsatian

3.1. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the task of automatically transcribing audio speech into written text. The field has achieved significant results in recent years with the emergence of multilingual models, such as Hubert Hsu et al. (2021), or Chen et al. (2022). Models such as Keren et al. (2025) claim to be able to transcribe thousands of different languages. Whisper Radford et al. (2023) is able to transcribe over 50 languages out of the box, although with significant differences in quality. These new multilingual models are able to produce good results on languages with smaller training datasets, by taking advantage of larger training data in other languages. They also allow finetuning : re-training a model with a small amount of data from a new language. Our previous experiments on Whipser [4] have found that as little as 30 hours can be enough to drop the Word Error Rate (WER) from around 75% to around 10%, depending on source and target language. However, these experiments have also shown that absence of an orthographic standard, which is the case in Alsatian, is an important limiting factor.

3.2. Existing datasets

Currently available voice datasets for Alsatian come from either broadcast recordings, or ethno-linguistic studies.

The **Linguistic atlas of Alsace** Huck et al. (2014) is an ethno-linguistic corpus from 1965 - 1980 whose goal was to document the language of the time in its geographic variability. Its limitation are the age of the recordings, that might not represent the current language accurately, and be of lower quality ; and its large variability, an obstacle for machine learning.

FLARS is another ethno-linguistic corpus, from the project « Effects of the national border on the linguistic situation in the Upper Rhine area », gathered between 2012 and 2014 Erhart (2017). These recordings are more recent and thus better represent the current language, but still are focused on variability.

Broadcast archives are a potentially large source of data. However, getting access to this kind of data can be an arduous process, as they are typically under strong authorship protections. This type of recording also presents difficulties for ASR training : There can be background music, overlap of voices, and it can be difficult to identify all participants to get consent. They also rarely come with transcription, although seven of them were fully transcripted in Erhart (2012). For these reasons, this type of data is not included in this study.

None of those sources of data are made for machine learning, and share this same limitation : Audio files are several minutes long, if not hours, and if there is a transcription, it is not aligned at a smaller granularity. These datasets must go through the preprocessing step of chunking them into smaller parts, and align the transcription to these chunks, which is costly.

3.3. ASR experiments

These experiments were performed with Whisper, on whisper-large-v3 model constrained to German. We chose German because it is the closest available language to Alsatian, and is one of the languages where Whisper performs the best, with a WER of 5.7%.

Whisper at finetuning requires segments of 30 seconds at most. To chunk both audio and transcription, we used Montreal Forced Aligner McAuliffe et al. (2017).

The results of the fine-tuning experiment are presented in table 1. As expected, results on the German ASR model with no finetuning are very low. Finetuning with the few hours of data at our disposal raises significantly the quality of the model, showing that even a small increase in available data can be useful. But current available data is not enough to bring the results to an acceptable level.

Following these results, we decided to launch a crowdsourcing campaign. This will allow to generate a new dataset that will be bigger, already chunked into small segments, transcribed, and devoid of music or other noise.

Name	Dataset size	WER	CER
No finenuning	-	90.66	57.22
FLARS	9h 11min	48.89	27.47
Atlas	2h 42min	58.70	33.48

Table 1: Results of fine-tuning. Results are expressed in Word Error Rate and Character Error Rate as a percentage. Results are evaluated on FLARS and Atlas data.

[4] citation redacted for anonymisation

4. Crowdsourcing campaign

4.1. Preparation steps

Common Voice Ardila et al. (2020) is a platform that uses crowdsourcing to create voice data in a large panel of languages. It became one of the standard benchmarks for automatic speech recognition. In July 2025, it launched a new mode : Spontaneous Speech. Instead of reading a pre-written sentence, participants can freely answer a given question or prompt. This new mode reduces the amount of work needed to launch a new language : instead of entering hundred of written sentences for participants to read aloud (which might be challenging to gather in a predominantly oral language), only 60 prompts are needed to launch a language. Since participants are free to answer those prompts in their own words, the resulting dataset contains a larger variety of sentences. It also allows participants to contribute in their own varieties of their language, as long as the inter-comprehension is sufficient to understand the prompt.

We created 60 prompts trying to strike a balance in representing the different varieties of Alsatian, while keeping the sentences understandable for a speaker of any variety. The Alsatian adaptation could have been restricted to that, given that users would have the option to use the platform in their preferred language (the platform being already available in French, German or English). However, to avoid confusing participants and creating language interferences, and given that one of the campaign's objectives was to demonstrate to Alsatian speakers that their dialect can be used in any situation, regardless of its written form, it was decided to adapt the entire website into Alsatian. As Alsatian is not included in international classifications and does not have its own ISO code, it is listed under the abbreviation "gsw", which actually stands for Swiss German.

The translation was undertaken by three authors of this paper. All three are linguists and have Alsatian as their first language. The translation process was challenging due the characteristics of Alsatian described above. First, since only one Alsatian version of Common Voice could be offered, the decision had to be made as to in which Alsatian variety the site would be translated. The Low Alemannic dialect was selected given its prevalence (still with many variations) in 90 per cent of the Alsatian region. To make the platform as inclusive as possible of all dialect varieties, we made exceptions in cases where the lexical differences between varieties were too significant. For example, the verb "listen" can have two different forms: "horiche" in the northern and main part of Alsace and "loose" in the extreme south. We therefore decided to include both forms in the category appearing on the platform, i.e. "horiche/loose".

In addition, the issue of language contact effects had to be addressed. As a dialect of the High German group, Alsatian is spoken by individuals who also speak French and also have contact, at different levels, with German and/or English. A parallel lexicon of the four languages was established based on the three existing versions of the Common Voice website (English, German, French). The Alsatian form for each entry was chosen on the following criteria: In cases where a specific term existed in Alsatian, it was retained systematically, provided that it appeared sufficiently widespread and transferable within the context of the platform's usage. Otherwise, it was necessary to assess to what extent a transposition from Standard German could be understandable/acceptable, and whether it would be better to borrow the word from French or English, and if so, the extent to which it should be integrated into Alsatian. For instance, the Alsatian verb "fùrtschìcke" was chosen over the Standard German "absenden" (which could have been adapted into "àbsände"), and "Gschirrkischt" was selected over the English "toolbox", which was chosen for the German version of the website. Although the referent of the Alsatian word is usually a real object, we believe that it can also be used in an abstract sense specific to the digital world, and that this unconventional use may appeal to users. In comparison, we decided not to translate the English term "cookies", despite the existence of an equivalent Alsatian term ("bredle"), as the latter is too specific. It refers to Alsatian Christmas cakes and has entered regional French usage, meaning any other use would undoubtedly seem inappropriate.

The management of internationalisms and terminology specific to digital usage has also led to specific terminological choices in Alsatian relating to neology Erhart (2023a). The term "avatar" is used in French, German and English, for example, but since its use is not attested in Alsatian, it was decided to propose a term that would be more meaningful to users, namely "Profilbild," literally "profile picture". While many English terms in use in the digital world were not translated into German, it was decided to translate them directly into Alsatian, basing on the historical relationship between the two languages stemming from common Old Germanic : for example, the verbs to "upload" and "to download" were translated into the neologisms "hochlàde" and "runterlàde". The decision was also taken to translate the English word "webbrowser" by creating a new term, "Webbrüser". This was motivated by the wish

to illustrate the possibility of lexical innovation in Alsatian, and was based on a play with the sounds of the English word ”browser” and the Alsatian noun ”brüser”, which is derived from the verb ”brüse” (German ”brausen”) meaning ”to produce a steady, uninterrupted noise”.

Finally, the question of the graphic choices made for this translation arose with particular sharpness: the Alsatian version of the site had to be clearly distinct from both the German and French versions, while remaining readable for the majority of users who speak the different varieties of Alsatian and are not used to read or write it. Consequently, a relatively flexible approach to the ORTHAL method (see above) was adopted, using diacritical marks only for vowels characterised by specific pronunciation in Alsatian: à, ì, ù.

For all of the translation choices made, an attempt was made to achieve a balance between predictability, acceptability, and the creativity of the speakers ; these three criteria had previously been suggested by LePage and Tabouret-Keller (1985) to characterize linguistically heterogeneous situations, such as the Alsatian one. From this point of view, it would be interesting to know whether participants in the collection use the interface in Alsatian or in other languages and, if so, whether or not they adopt the terminology and/or spellings proposed in the Alsatian version of the site.

4.2. Results and feedback from the public

Events : A crowdsourcing campaign is only as strong as its communication plan. To generate interest, we held two in-person events : A press conference, presented as the launch of the campaign in October 2025, and a conference for the general public, in January 2026. Many of the exchanges we report below happened during one of those events, or in conversations immediately following them.

Reception and collection results : The press conference resulted in 10 press articles in national regional newspapers and radios. These included articles in two major daily regional newspapers, *L'Alsace* and *Dernières Nouvelles d'Alsace*, as well as the morning programme on the main regional public radio station, *Ici Alsace*. As of the writing of this paper, we collected 5 hours and 34 minutes of voice data. However, only 35 minutes have been transcribed.

Reactions regarding Alsatian and AI The way in which the Common Voice Spontaneous Speech data collection for Alsatian was presented in the press provides an initial indication of how it was perceived by journalists. For example, one of them admitted that he didn't really understand the purpose of the data collection. In his article, he wrote that « the language used by researchers can sometimes be obscure », yet he still invited Alsatian speakers to « donate their voice » and « help researchers develop regional languages, such as Alsatian, in the digital sphere ». Another article was published in the ”Insolite” (bizarre, quirky) section of the *Ici Alsace* radio station's website. This may suggest that the project was not taken seriously, or at least that it was perceived as an attempt to bring together two seemingly incompatible worlds: the Alsatian language and computational sciences.

This brings us to what appears to be the main obstacle to encouraging people to participate in this spectific data collection: the ideologies and attitudes of speakers towards their languages, and towards language technologies in general. Indeed, the reactions observed during the organised events, the messages received as well as the comments on press articles shared on social media generally confirm the common perception of Alsatian as a language linked to the past and with very low social prestige. These perceptions also suggest that its typical form is spoken by elderly people who are rural and less mobile, who have less contact with other languages and who are the usual witnesses of traditional dialect geolinguistics Bothorel-Witz (2007).

Consequently, the latter are regarded by the general public as the ideal candidates for the collection (one student told us « I will tell my grandpa about it » ; a colleague suggested going to a mountain cabin he visited on vacation where he heard the owners speaking Alsatian), whereas the project aims, in fact, to reach all speakers of Alsatian and to collect and document Alsatian as it is spoken today, encompassing its entire diversity. However, elderly speakers are less likely to be comfortable with digital tools and with participating in such a data collection exercise without the presence of an interviewer who would ask them questions in person, like in traditional dialect data gatherings.

Furthermore, variations in the language are subject to contradictory representations; they are perceived both as a defining characteristic of Alsatian and as an obstacle to mutual understanding between its different varieties. It has been repeatedly expressed as a concern that the Common Voice platform and the questions asked for the data collection do not correspond to the ”Alsatian” spoken by the participants.

Finally, even though the platform is primarily aimed at collecting voices in Alsatian, it confronts its users with written Alsatian, which they are not used to. Indeed, according to Huck (2002), in the perception of society, written Alsatian belongs to the « realm of the anecdotal and incidental, a realm

it has never really left since its emergence in the nineteenth century ». Despite its well-established use in digital communication Erhart (2025) Erhart (2026), written Alsatian remains largely used for informal communication. More formal uses, such as the tasks to be completed on Common Voice, i.e. answering questions formulated in writing in different varieties of the language or transcribing audio recordings in Alsatian into writing, are not yet widespread. While reading the questions does not appear to be an insurmountable task, it is noteworthy that most of the participants, probably concerned about their ability to transcribe Alsatian correctly, give up on the transcription activity. This would explain the low volume of transcribed data collected to date.

Reactions regarding technology and the campaign's goal

Eventually, to understand why people do or do not decide to participate in the collection, it is necessary to integrate the analysis of their attitudes towards the Alsatian language with that of their attitudes towards technology and their comprehension of the campaign's objective.

The press campaign's association of Alsatian with Artificial Intelligence (AI) may have resulted in potential candidates experiencing confusion, as many of them tend to associate AI with generative AI and anticipate (or fear) the emergence of an "Alsatian ChatGPT." Concerns were also expressed regarding the utilisation of the collected data and the ethical and personal data protection issues that would arise. (« So *anyone* can use my recordings to do *anything*? »)

There was also confusion between the activities of researchers and activities of the industry. If we explain that the campaign will help develop ASR technologies, people tend to understand that we will personally develop the application that will appear by itself on their phone. If we say we won't, they become confused about what it is we do, exactly, which might even erode trust into public research.

All of these examples are obstacles to public interest and involvement. To circumvent these obstacles, effective communication is imperative. In our experience, the most effective method of overcoming such potential resistances is through personalised, face-to-face or at least interpersonal explanations.

4.3. Technical difficulties

Finally, in part because Alsatian has been pigeonholed into a topic of interest for the older generation, we had many users confused about how to use the platform. Common Voice is an old and large project. Users had trouble navigating the interface of the large website to find the Spontaneous Speech section. Some participants did not understand that the task consisted in answering the questions and recorded themselves reading the questions out loud. In some cases, their browser automatically attempted to translate the Alsatian prompts into French, a functionality that the users didn't know existed nor how to turn it off.

We mitigated these issues by creating a homepage giving out simple instructions and direct links to the relevant pages, and asked the press to share a link to this homepage, rather than directly Common Voice. We gave a contact e-mail in all our communications, and were reactive in answering many requests for support, with kind help from Mozilla when we couldn't solve an issue ourselves. However, it was difficult to help people navigate technical difficulties remotely. This is why we organised the second, public, event, and included time for the participants to set up their account while we were physically in the room and able to assist them.

From this, we conclude that there are advantages and issues with using a large, well-recognized platform such as Common Voice, rather than creating your own platform for your usage : The platform is more complex, less easy to navigate for your users. But the advantage is the technical maturity of the platform (capturing voice from many different devices and browsers is not trivial) and the visibility and perenity of the dataset.

4.4. Feedback from the data collection on participants motivations

To gain insight into participants' motivations, the following specific question was included in the first 60 Alsatian prompts of the Spontaneous Speech mode of Common Voice : « Wàrùm màche-n-r dann àn Common Voice mìt? » (Why do you participate in Common Voice ?). At the time of writing this article, 19 out of the 47 participants in the collection had recorded a response to this question.

The decline in the use of Alsatian is explicitly stated by several of them, who also express their regret about it (« ùn dìss ìsch Schàd » : and it's a shame). While most of the respondents consider Alsatian and/or its preservation to be « important » (the adjective « wichtig » was used seven times), it is interesting to observe that their attitudes are complex and ambivalent, and may diverge in some respects. The motivation behind their participation appears to be twofold: firstly, the desire to circumvent a potential loss, and secondly, the aspiration to perpetuate a linguistic tradition that has endured for several centuries, Alsatian having served as the primary medium of intergenerational communication in the region from the Middle Ages until the advent of the 20th century.

The choice of verbs used in the responses

reveals slightly different issues: the use of the verbs *verschwinde* (to disappear), *vergasse* (to forget) and *verliere* (to lose), preceded by the negative adverb *nitt*, indicates above all a desire to preserve the language before it disappears, while the use of verbs such as *bhàlte* (to keep), *preserviera* (to preserve) and *bliewe* (to stay) indicates more of a desire to preserve what one participant compares to a treasure (*Schàtz*). But what seems to be at stake in most of the responses is the transmission of Alsatian to future generations (*nächste Generatione*), as reflected by the frequency of verbs formed with the adverb *widderscht* (further) such as *widderscht gehn* (to go on), *widderscht redde* (to keep speaking) and *widderscht gan* (to pass on). Some participants even consider that the survival of the Alsatian language is at stake, as shown by the use of compound verbs based on the verb *lawe* (to live) : one uses the verb *widderschtlabe* (to keep leaving) while others use the verb *ìwwerläwe* (to survive), as in the following response : « duich de Common Voice kànn s Elsässische vielleicht iwwerläwe » (thanks to Common Voice, the Alsatian language can perhaps survive). However, it is also worth noting the recurring use of modalizers such as *vielleicht* (perhaps), which serve to relativise the hopes expressed.

The importance of the participants' emotional responses and personal commitment is emphasised by the frequent use of possessive adjectives. Three participants use a possessive adjective in the first person singular before the term "mother tongue", which they use to refer to Alsatian (e.g. « ich will zeje wie mini Muetersprooch isch »: I want to show what my mother tongue is), while six other participants use a possessive adjective in the plural before another name for the language (e.g. « fer dàss mr unser Dialekt euh bhàlte kànn »: so that we can keep our dialect), the referent to which the plural refers remaining implicit. With regard to the way they name the Alsatian language, it is noteworthy that only one participant used the term "dialect" to refer to Alsatian, with the majority using *Elsassischi Sproch* or *Elsassisch*. While one participant expressed a strong personal commitment to the preservation of their language, by saying « ich màch, àlles wàs ich kann » (I do everything I can), several other express their willingness to contribute in their own humble way (« wenn ich e bìssel helfe kànn »: if I can help a little). One participant's involvement in the collection appears to serve as a form of compensation for a sense of guilt surrounding the non-transmission of Alsatian, as articulated in his testimony (« do sin mir Eltere d erschte d Schùld dràn »: we, the parents, are responsible for it).

Regardless of the positions expressed by each individual participant, it appears clearly that the motivation behind their participation in the collection was the preservation of the language, rather than a interest in new technologies. This is a crucial aspect that must be considered in future communications related to the Alsatian Spontaneaous Speech data collection.

5. Conclusion

In this study, we launched an Alsatian version of Common Voice to address a lack of voice data in sufficient quantity and quality. Our campaign gathered, at the time of writing, 5 hours and 34 minutes of data, although only 35 minutes have been transcribed. Future work will be focused on transcribing this data, and evaluating the impact of this new data on finetuning an ASR model.

By launching an Alsatian version of Common Voice, we gave the opportunity to Alsatian speakers to play an active role in the preservation of their linguistic heritage within the context of the modern digital world. However, the fact that less than 50 participants took part in the crowdsourcing campaign seems to indicate a lack of interest in the general public in this issue. As has already been shown in previous studies Erhart (2023b), a proportion of Alsatian speakers appear to be resigned to its potential extinction. This suggests a discrepancy between society's evident interest in this language and the actual willingness of speakers to revive its use. It is conceivable that Alsatian has become a post-vernacular language Shandler (2006) in the sense that its importance lies now more in its representation than in its actual use.

In any case, the implementation of this campaign revealed sociolinguistic issues that we had not sufficiently anticipated, in particular a significant gap between the current uses and representations of Alsatian and the use of language technologies, leading to what could be called a « digital code-switching ». This makes us question our role as researchers in the potential revitalization of the languages we work on. If we want to influence speakers' practices and ideologies about languages, and encourage participation in campaigns like the one we launched, effective communication about the research issues is key. We consider that the preservation of linguistic diversity on a global scale is at stake, along with « the possibility to think differently, to produce worlds that do not merely reproduce, in impoverished forms, the codes of a hegemonic culture, but which embody the vitality of multiple cosmologies, to speak languages that are not mere means, but universes, to design technologies that do not standardise the world, but expand it » Baraldi and

Rico Menge (2025).

6. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus.

Luca Baraldi and Ma. del Carmen Rico Menge. 2025. Vers une nouvelle épistémologie : repenser le pouvoir, la technologie et le langage. *Cahiers du plurilinguisme européen*, 17.

Arlette Bothorel-Witz. 2007. L'Alsace et ses langues. Eléments de description d'une situation sociolinguistique en zone frontalière. Variétés en contact et représentations sociolinguistique [sic]. In *Aspects of Multilingualism in European Border Regions. Insights and Views from Alsace, Eastern Macedonia and Thrace, Lublin Voivodeship and South Tyrol*, pages 39–56. EURAC Research (Europäische Akademie) .

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Collectivité européenne d'Alsace. 2022. Étude sociolinguistique sur l'alsacien et l'allemand. Rapport de présentation, mai 2022 .

Pascale Erhart. 2012. *Les dialectes dans les médias : quelle image de l'Alsace véhiculent-ils dans les émissions de la télévision régionale ?* Theses, Université de Strasbourg.

Pascale Erhart. 2017. Les effets de la frontière sur les pratiques linguistiques dans le Rhin Supérieur . *Cahiers du plurilinguisme européen*, (9).

Pascale Erhart. 2023a. « Hett dis kenn nàmme uf elsässisch ? » La néologie en alsacien par le prisme des émissions de la télévision régionale. *Neologica : revue internationale de la néologie*, (17):47–72.

Pascale Erhart. 2023b. L'alsacien ne vaut rien, mais rien ne vaut l'alsacien ! In *Appartenances, marchés et mobilités : penser la valeur des langues*, Sociolinguistique, pages 71–80. L'Harmattan.

Pascale Erhart. 2025. « One does not simply bstell e Flammküeche ohne Ziwwle »: Sociolinguistic Issues of Multilingual Computer Mediated Communication in Alsace. In *Social Media Current Issues in Romance Linguistics and Foreign Language Education*, pages 117–145. AVM.edition.

Pascale Erhart. 2026. Elsässisch 2.0. Dialektgebrauch in der digitalen Kommunikation am Anfang des 21. Jahrhunderts im Elsass. In *Dialekt in Gesellschaft und Schule. Variation und Wandel in Gebrauch und Wahrnehmung des Alemannischen*, pages 103–124. Franz Steiner Verlag.

Lucile Hamm. 2024. Enseignement de l'alsacien / en alsacien à l'école maternelle publique : (nouvelles) perspectives de recherche sur le terrain alsacien ? *Synergies Pays germanophones*, (17).

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. glottolog/glottolog: Glottolog database 5.2.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Dominique Huck. 2002. Les dialectes en Alsace : fonctions et statut de l'écrit dialectal et normes graphiques. Etat d'un non-débat. In *Codification des langues de France. Actes du colloque « Les langues de France et leur codification »*, pages 99–110, Paris, Inalco.

Dominique Huck, Arlette Bothorel-Witz, Sylviane Spindler, Ernest Beyer, Raymond Matzen, and Marthe Philipp. 2014. Atlas Linguistique et ethnographique de l'Alsace.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. ArXiv preprint arXiv:2511.09690.

Anaïs Koehler. 2024. *L'impact du français sur les parlers dialectaux des locuteurs de l'Alsace contemporaine*. Theses, Université de Strasbourg.

Robert B. Le Page and Andrée Tabouret-Keller. 1985. *Acts of identity: Creole-based approaches to language and ethnicity*. Cambridge University Press.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *40th International Conference on Machine Learning*, pages 28492–28518.

Jeffrey Shandler. 2006. *Adventures in Yiddishland. Postvernacular Language and Culture*. University of California Press.

Edgar Zeidler and Danielle Crévenat-Werner. 2008. *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*. Do Bentzinger.

HeptaTAX: A Neuro-Symbolic Pipeline and Benchmark for Classifying 16th-Century Heptanesian Notarial Acts

Stergios Chatzikyriakidis[1], **Eleni Karantzola**[2], **Vasiliki Makri**[3]
[1] University of Crete
[2] University of the Aegean
[3] National and Kapodistrian University of Athens
stergios.chatzikyriakidis@uoc.gr, karantzola@aegean.gr, vasmakri@enl.uoa.gr

Abstract

This study originates in the investigation of lexical bundles and formulaic language within sixteenth-century Corfiot notarial documents. The observed functional variation across identical formulaic sequences motivated the development of a document classification framework designed to support the structural interpretation of such language. Given that 16th-century Corfiot notarial acts represent a rich, albeit understudied, dialectal resource, their systematic categorization into subgenres is essential for their full exploration. However, this task requires substantial manual work, while NLP tools for this task and dialect do not exist. In this paper, we attempt to take an initial step in this direction. First, we present a corpus of 1,088 notarial acts from 5 notaries spanning 1500-1567, a 3-tier annotation schema (17 core genres, extension subcategories, hybrid cross-cutting tags), and a 40-act benchmark with gold annotations at all three tiers. Then, we evaluate 12 LLMs across 4 architectures, zero-shot, few-shot, full-context and Neuro-Symbolic. For the latter, we introduce a symbolic engine comprising a set of deterministic rules for identifying discriminative legal formulae, whose output is then injected into the neural (LLM) engine. The results show that the NeSy architecture compresses the accuracy gap between stronger and weaker models from 47.5 pp to 12.5 pp, with the smallest model (Llama 3.1 8B) gaining 47.5% and matching frontier models that operate without symbolic support. Three models reach a ceiling of 72.5% on the core tier. However, consistent errors in procedurally dense material reveal the limits of lexical and formulaic cues for identifying legal effect, motivating the use of symbolic signals in the NeSy pipeline. Extension and hybrid classification remain open challenges, with best scores of $\sim$63% and $\sim$35% respectively.

Keywords: NLP, neuro-symbolic, Heptanesian Greek, Corfiot, notarial acts, document taxonomy

1. Introduction

The sixteenth century marks a crucial phase for the study of post-Byzantine vernacular / Early Modern Greek, characterized by extensive linguistic variation and the absence of a standardized norm (Karantzola, 2024; Kakoulidi-Panou et al., 2023). Within this context, notarial documents from Corfu provide a uniquely valuable corpus: they are precisely dated, geographically localized, and shaped by recurring formulae that capture the interaction between professional writing practices and regional linguistic features. Classifying these notarial acts into legal genres (Venditio, Testamentum, Emphyteusis, etc.) is essential both for historical interpretation and for the linguistic analysis of formulaic language. However, this process requires domain expertise, remains extremely time-consuming when performed manually, and is further complicated by editorial inconsistency: not all editors classify documents in the same way, and proposed categorizations may be guided by surface lexical cues rather than the legal effect of the instrument. On the other hand, performing this task computationally using NLP techniques is not straightforward either, since systems have to face non-standard orthography, code-switching (Greek/Latin/Italian), formulaic but variable legal language, and domain vocabulary absent from LLM training data. We present a first attempt to tackle this problem via HeptaTAX offering the following three research contributions:

1. A new corpus of $\sim$1,088 notarial acts from 5 sixteenth-century Heptanesian[1] and in particular Corfiot notaries along with a 40-act benchmark including 3-tier genre annotations.
2. A taxonomy of 17 core legal genres with extensions and hybrid tags based on the close reading of notarial acts by experts.
3. A neuro-symbolic pipeline that combines deterministic symbolic rules with LLMs.

We then evaluate 12 LLMs (both open and close weight ones) across 4 architectures (zero-shot, few-shot, full context, and NeSy), effectively showing that symbolic knowledge injection can help in reducing the gap in performance between stronger

[1] Despite its acknowledged contribution to the formation of Standard Modern Greek, Heptanesian Greek, which is spoken in the Ionian Islands (as well as Kythira), forms a distinct and cohesive dialectal group characterized by a set of grammatical and lexical features that set them apart both from the standard language and from mainland dialects; internal variation among the islands further supports their independent status (Liosis, 2024).

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 265–273
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and weaker models.[2]

2. Related Work

Within the Early Modern period (16th–18th centuries), notarial acts constitute a key linguistic source for the study of vernacular Greek, as their precise dating, geographical localization, and formulaic structure enable reliable contextualization and systematic corpus-based analysis (Karantzola et al., 2012; Manolessou, 2003; Makri, 2020; Wagner et al., 2013). The classification of notarial documents into documentary types has long been a central concern of diplomatics, where distinctions such as *instrumenta* and *acta* reflect functional differentiation in written practice (Gasparini, 2023; Duranti, 2015). In digital corpora, document typology constitutes essential metadata for both historical interpretation and computational processing, motivating recent work on the automatic classification of historical documents (Ehrmann et al., 2023). At the same time, the strong formulaicity of notarial writing has prompted linguistic research on recurring lexical bundles and formulaic sequences (Biber et al., 2004; Wray, 2002).

NLP work for Greek dialects is not abundant but has seen a surge in recent years. For example, Chatzikyriakidis et al. (2023) introduce GRDD, a large-scale raw text dataset for Greek dialectal NLP that covers four MG dialects, i.e., Cretan, Pontic, Northern Greek, and Cypriot. This has developed into GRDD+ (Chatzikyriakidis et al., 2026a), further extending the resource to 10 varieties (6.3M words), including Heptanesian, and also performing fine-tuning experiments on Llama and Krikri models on four dialects. There is also important work on treebanks for Greek dialects: Eastern Cretan (Vakirtzian et al., 2025) and the dialect of Lesbos (Bompolas et al., 2025), both using cross-dialectal knowledge transfer from Standard Modern Greek. Vakirtzian et al. (2024) also introduced a speech recognition benchmark for four Greek varieties. Chatzikyriakidis et al. (2026b) apply perplexity as a metric for measuring dialectal distance across Greek varieties. Earlier attempts to provide computational resources include the work by Anastasopoulos et al. (2018), who present a POS-tagged dataset for Griko, and the Multi-CAST version for Cypriot Greek by Hadjidas and Vollmer (2015). There is also dialectometric work for Greek dialects. There, we find the work by Bompolas (2023) that applies traditional dialectometric distance-based methods to Asia Minor varieties (Cappadocian, Pharasiot, and Silliot) and also Psaltaki et al. (2025) that deals with loanword detection of Turkish and Italian in four dialects of Modern Greek. These are very important initiatives to develop resources and models for Greek dialectal NLP, but none of them deal with the classification of historical documents, let alone documents in 16th-century Heptanesian.

On the neuro-symbolic (NeSy) side, our system is a Federative-Reasoning (F-R) NeSy system under the Chatzikyriakidis-Lappin taxonomy (Chatzikyriakidis and Lappin, To appear), i.e. one in which the symbolic and neural components operate as fully autonomous modules that communicate only through their outputs, without any modification of the internals of either component. Similar systems have been proposed for various NLP tasks, but a detailed discussion of such systems falls outside the scope of this paper. However, there is a closely related F-R system that uses a hybrid phonological system to deal with the identification and generation of Greek rhyme (Chatzikyriakidis and Natsina, 2026). There, a deterministic rhyme-detection module verifies and refines LLM-generated Modern Greek poetry, raising accuracy from 4% to 73%. The symbolic component in the system described in this paper plays an analogous role: it preprocesses the input and passes structured evidence to the LLM, which retains the freedom to override it.

3. The HeptaTAX Resource

3.1. Corpus

The corpus comprises five published registers of Corfiot notaries, originally edited from manuscript sources and digitized at the Laboratory of Southeastern Mediterranean Linguistics (University of the Aegean). The texts were processed via OCR, normalized to monotonic orthography, and verified against the editions, including manual correction and even typing where required. The corpus contains 1088 acts, distributed per notary as follows:

- Alexakis: 104 acts (1513–1516)
- Farmakis: 173 acts (1515–1525)
- Spyris[3]: 375 acts (1560–1567)
- Toxotis: 282 acts (1500–1503)
- Varagas: 154 acts (1541–1545)

The sixteenth-century Corfiot notarial documents reflect an early stage in the formation of the Heptanesian varieties and provide a valuable source for historical dialectology despite the constraints of

[2]The data and the code needed to reproduce the results can be found here: https://osf.io/vj2fr/overview?view_only=a9cccb01cae148bd9c095e50d687a261

[3]It should be noted that a subset of 19 acts was authored by the notary Photios Palatianos.

their legal-administrative register. Although historical orthography may obscure phonological developments in progress, the texts combine conservative features with emerging morphophonological, morphological, and lexical innovations, indicating an ongoing phase of linguistic reorganization rather than a stabilized dialectal system (Karantzola and Lavidas, 2016; Karantzola and Makri, in print). Variation across scribes and communities points to both diatopic (geographical) and diastratic (sociolinguistic) differentiation, while pervasive orthographic fluctuation reflects the absence of a fixed written norm. Overall, the corpus occupies an intermediate position between administrative convention and vernacular usage, preserving regional traits while documenting broader processes of linguistic change.

3.2. Benchmark

In order to test the ability of NLP models to classify the genre of notarial acts, we sample 40 acts from the corpus, corresponding to 8 acts per notary. Two experts then annotate them using a 3-tier label scheme that includes:

- Core layer: one of 17 legal genre categories (Table 1)
- Extension layer: subcategory where applicable (e.g., Venditio → Venditio cum pacto retrovendendi)
- Hybrid tags: cross-cutting features (e.g., "cum pacto", "agricultural credit")

Annotation was performed collaboratively by two domain experts, who jointly assigned labels through discussion until consensus was reached.

3.3. Taxonomy

The two experts also built a taxonomy of 17 core legal genre categories and organized them into 6 thematic groups (Duranti, 2015; Fellmeth and Horwitz, 2021):

I. *Contractus de rebus* (property): Venditio, Donatio, Permutatio, Emphyteusis, Locatio conductio rei

II. *Contractus de credito* (credit): Mutuum, Instrumentum obligationis, Quietantia, Pignus/Hypotheca

III. *Contractus operarum* (production): Locatio conductio operis, Societas, Praevenditio

IV. *Acta familiaria* (family/personal): Testamentum, Dos, Divisio/Partitio

V. *Acta proceduralia* (procedural): Procuratio, Renuntiatio, Revocatio

VI. *Acta ecclesiastica*: Concessio ecclesiastica, Beneficium

The Extension layer then captures subtypes of the main taxonomy (e.g., emphyteusis perpetua, locatio vineae), while the hybrid layer captures cross-cutting features that span multiple core types. The taxonomy is presented in trilingual form: Latin / English / Greek.

4. Methodology

We evaluate four prompting strategies across 12 LLMs. The first one is a **zero-shot** setting, where the model receives only the act text together with the list of 17 valid core categories and an instruction to classify. The second is a **few-shot** setting, where the prompt is augmented with 5 manually annotated examples covering diverse genres (Venditio, Praevenditio, Emphyteusis, Procedural acta, and Concessio ecclesiastica), each accompanied by its gold 3-tier label. The third is a **full-context** setting, where the model additionally receives the complete genre taxonomy manual, including definitions of all 17 core categories, extension subcategories, and hybrid tags. The fourth strategy is a **neuro-symbolic (NeSy)** setting. Here, we pair the full-context prompt with the output of a deterministic symbolic component. This component comprises ∼70 hand-crafted rules that pattern-match on formulaic phrases in the normalised act text. The symbolic rules map a phrase to a genre and carry a weight, ranging from 3 (near-unique to one genre) to 1 (weakly associated). Any matched rule is collected and are then used to rank the candidate genres by cumulative weight, assigning, at the same time, a confidence level based on the top score and the margin over the runner-up. The result is a symmary that includes the predicted genre, confidence, and supporting phrases, that is passed to the LLM with the instruction to treat this information as defeasible. It is important to note that the symbolic component runs locally before the API call and, as such, imposes no extra inference cost.

The NeSy system proposed is an instance of an **Federative-Reasoning (F-R)** architecture in the Chatzikyriakidis-Lappin taxonomy (Chatzikyriakidis and Lappin, To appear), where the symbolic engine is totally autonomous from the neural component. The idea is that the output of the symbolic engine is passed to the neural component (the LLM in our case) without the involvement of any modification of the internals of each component. The motivation for using a NeSy system in our case is quite standard: LLMs generalise well across a remarkable number of tasks but sometimes fail because they lack domain-specific knowledge, while, on the other hand, hand-crafted rules encode precise formulaic

Group	Core Genres	N
I. Contractus de rebus (property)	Venditio, Donatio, Permutatio, Emphyteusis, Locatio conductio rei	5
II. Contractus de credito (credit)	Mutuum, Instrumentum obligationis, Quietantia, Pignus/Hypotheca	4
III. Contractus operarum (production)	Locatio conductio operis, Societas, Praevenditio	3
IV. Acta familiaria (family/personal)	Testamentum, Dos, Divisio/Partitio	3
V. Acta proceduralia (procedural)	Procuratio, Renuntiatio, Revocatio	3
VI. Acta ecclesiastica	Concessio ecclesiastica, Beneficium	2
	Total	**17**

Table 1: The 17 core legal genres organised in 6 thematic groups following the Latin legal tradition adapted for Venetian-ruled Greek practice. The extension layer captures subtypes (e.g., *emphyteusis perpetua*, *locatio vineae*); the hybrid layer captures cross-cutting features that span multiple core types.

Normalised phrase	Genre	W
ψιχικις μοι σοτιριας *'for my soul's salvation'*	Testamentum	3
εποιλισεν *'sold'*	Venditio	3
χανονιν *'canon / annual rent'*	Emphyteusis	3
πλιρομενος χε ειχαριστιμενος *'paid and satisfied'*	Quietantia	3
ις τιν ερχομενιν εσοδιαν *'in the coming harvest'*	Praevenditio	3
χριτε αλπιτρε *'judges arbiters'*	Procedural acta	3
δια τιμιν *'for a price'*	Venditio	2
επαρεδοσεν *'delivered' (generic transfer)*	Venditio	1

Table 2: Sample symbolic rules. Each rule maps a normalised formulaic phrase to a target genre with a discriminative weight (W): 3 = near-unique to one genre, 2 = strongly associated, 1 = weakly associated (shared across genres). The full rule set contains ∼70 rules.

knowledge but break easily when faced with open text (Ebrahimi et al., 2024).

Table 2 presents a sample of the symbolic rules, illustrating how genre-diagnostic formulaic phrases from the notarial tradition are mapped to target genres with varying discriminative weights.

All 12 models ran through the OpenRouter API at temperature 0.0. We have a good selection of models that cover both proprietary and open models but also of different sizes and architectures (thinking vs non-thinking). Most specifically we use 6 proprietary models (Claude 3.7 Sonnet, Claude 4 Sonnet, Claude 4 Opus, GPT-4o, GPT-5.2 Chat, and Gemini-2.5-Flash) and six open-weight models (Llama 3.3 70B, Llama 4 Maverick, DeepSeek V3, DeepSeek R1, Qwen 2.5 72B, and Llama 3.1 8B). Each model is asked to produce a structured JSON response with three fields as output: `core_layer`, `extension_layer`, and `hybrid_tags`. In total, 12 models × 4 strategies × 40 acts = 1,920 classifications.

5. Results and Analysis

The results are summarized in two tables: Table 3 reports the accuracy for the core layer only, while Table 4 reports the results for the 3-tier evaluation (core, extension, hybrid). We first discuss the non-symbolic strategies before turning to the NeSy results.

Zero-shot and few-shot. Starting with the zero-shot setting, models range from 17.5% (Llama 3.1 8B) to 65.0% (GPT-5.2 Chat and Claude 3.7 Sonnet), with a mean of 53.3%. The main factor here seems to be model scale. Smaller models (Qwen 2.5 72B, Llama 3.3 70B) as well as DeepSeek V3 score below 55%, while frontier models cluster in the 57.5–65% range. The addition of five annotated examples in the few-shot setting gets the mean up to 56.2%. However, the effect is not stable across models. Some models benefit considerably (DeepSeek V3: 45.0% → 57.5%; Qwen 2.5 72B: 45.0% → 55.0%), while for others it is detrimental to their performance (e.g. Llama 4 Maverick drops sharply from 60.0% to 37.5%, and GPT-4o from 62.5% to 57.5%). This suggests that the five examples may introduce a distributional bias that misleads some models.

Full-context. This is the case in which the models receive the whole taxonomy manual. Accuracy mean is raised to 58.1%. An interesting fact is that the increase is most visible in models that seem to struggle with the few-shot setting (Llama 4 Maverick recovers to 60.0%; Qwen 2.5 72B remains flat at 45.0%). However, the strongest gains are for the two Claude models used (Claude 3.7 Sonnet reaches 70.0%; Claude 4 Opus also reaches 70.0%). Full-context is the preferred strategy for Claude 4 Opus (70.0% vs 67.5% NeSy) and a competitive one for DeepSeek R1 (60.0%).

Neuro-symbolic (NeSy). The NeSy setting achieves the highest mean accuracy at 67.7% and is the best or tied-best strategy for 10 out of 12 models, with the two exceptions being Opus 4 and DeepSeek R1 as noted above. The top three NeSy models—GPT-4o, GPT-5.2 Chat, and Llama 4 Maverick—all reach 72.5%, followed by Claude 3.7 Sonnet and Claude 4 Sonnet at 70.0%. In the extension layer, the full-context and NeSy settings are close (best in 5/12 and 6/12 models, respectively). In the hybrid tags layer, accuracy remains low across the board (25–35%) with no clear winner among strategies. Notably, the NeSy setting considerably improves small models on extensions (Llama 8B: 15.4% full-context $\rightarrow$ 48.7% NeSy).

One of the main findings is that the NeSy approach acts as an equalizer for weaker and stronger models. For example, Llama 3.1 8B goes from an accuracy of 17.5% in the zero-shot setting to 65.0% in the NeSy setting (47.5-point increase), suggesting that good domain-specific rules can compensate for model scale.

There is an accuracy ceiling at 72.5%. Three architecturally different models (GPT-4o, GPT-5.2, Llama 4 Maverick) all plateau there. The remaining errors might point to task-inherent limitations. In particular, the performance ceiling we observe at 72.5% is likely due to the the inherent ambiguity of the historical legal register, where multiple juridical effects overlap within a single instrument, rather than from architectural constraints of the models.

DeepSeek-R1, the only model with mandatory chain-of-thought generation, is not helped by NeSy and achieves its best score in the few-shot setting. Moving to the frontier instruction-following models, results are mixed: Sonnet 3.7 and Opus 4 perform best in the full-context setting, while Sonnet 4 and GPT-5.2 benefit from the symbolic component.

Llama 3.1 8B in the NeSy setting (65.0%) presents a very interesting cost-efficiency ratio case. The 8B model is $\sim$150$\times$ and $\sim$750$\times$ cheaper per input and output token, respectively, than Claude Opus 4 (70% in full context setting, 67.5% in the NeSy setting). Proper NeSy architectures can make these much smaller models useful for similar tasks at a much lower cost.

Comparing full-context vs NeSy, we see evidence that providing the full taxonomy helps, but curated symbolic evidence on top of the full-context (NeSy) provides better performance in 10/12 models for the core layer. Moving on to the extension layer results, we find a slightly different picture. There, full-context settings outperform NeSy in 5/12 models (GPT-5.2: 62.9% vs 59.5%; Opus 4: 57.9% vs 48.7%; R1: 60.0% vs 51.4%). However, NeSy still improves the performance of small models on extensions. For example, Llama 8B goes from 15.4% (full-context) to 48.7% (NeSy), and LLama 3.3 70B. Lastly, the Hybrid tag remains challenging. Models struggle there within the 25-35% across all strategies, with no clear winner.

6. Error Analysis

Looking at patterns of confusion, we see that the errors cluster around categories with overlapping legal vocabulary. Examining the 48 attempts per act (12 models $\times$ 4 strategies), the most frequent confusions are: Procedural acta mistakenly labeled as Procuratio (47 cases), Locatio conductio rei as Emphyteusis (47 cases), Concessio ecclesiastica as Procuratio (46 cases), Procedural acta as Mutuum (46 cases), and Procuratio as Quietantia (44 cases).

Acts involving multiple legal transactions (e.g., sale combined with pledge) are among the hardest to classify, since the core layer forces a single label. It is also important to note that the symbolic rules do not cover a number of rare categories (e.g., Divisio or Beneficium), and as such, classification of these types relies entirely on the LLM's knowledge.

6.1. Procedural Registers as a Stress Test: The Toxotis Case

The Toxotis documents are different w.r.t the rest of the corpus. It is packed with procedural documents and relies heavily on institutional and delegation language. Unlike registers that mainly involve transactional acts, Toxotis' documents mostly deal with procedural maneuvering, for example they handle issues like appointments, declarations, setting up arbitration, or representing someone in a specific dispute.

Given this setup, the models, no matter the configuration, seem to assume they are looking at transactional documents, and as such, they greatly over-predict categories like Procuratio, property deals, and obligation-related classes. The results are quite telling: 5/8 Toxotis acts in the benchmark were missed by all 12 models across all 4 strategies (scoring 0 out of 48 attempts for each). The remaining 3 acts managed 2 correct predictions (out of 144 total attempts). The result is a 99.5% failure rate overall. It is important to note that these are not scattered mistakes, given that in most cases, all 12 models zero in on the exact same incorrect category, no matter the prompting strategy. For example, Act 130 (gold: Procedural acta / Declaratio) is classified as Mutuum by every model in every setting; Act 170 (gold: Procuratio) is unanimously labelled Quietantia; and Acts 156 and 137 are both called Procuratio by all 48 attempts despite their distinct gold labels (Procedural acta and Concessio ecclesiastica, respectively).

Misleading vocabulary appears to be the main

Model	Zero-shot	Few-shot	Full-context	NeSy
GPT-4o	62.5	57.5	62.5	**72.5**
GPT-5.2 Chat	65.0	70.0	70.0	**72.5**
Llama 4 Maverick	60.0	37.5	60.0	**72.5**
Claude 3.7 Sonnet	65.0	65.0	70.0	**70.0**
Claude 4 Sonnet	57.5	67.5	62.5	**70.0**
Claude 4 Opus	60.0	62.5	**70.0**	67.5
DeepSeek V3	45.0	57.5	55.0	**67.5**
Qwen 2.5 72B	45.0	55.0	45.0	**67.5**
Gemini 2.5 Flash	60.0	57.5	62.5	**65.0**
Llama 3.1 8B	17.5	30.0	22.5	**65.0**
Llama 3.3 70B	45.0	52.5	57.5	**62.5**
DeepSeek R1	57.5	**62.5**	60.0	60.0
Mean	53.3	56.2	58.1	**67.7**
Range (max–min)	47.5	40.0	47.5	**12.5**

Table 3: Core layer accuracy (%) across 12 LLMs and 4 prompting strategies on the 40-act benchmark. Bold indicates best strategy per model. NeSy = neuro-symbolic pipeline with symbolic evidence injection.

	Zero-shot			Few-shot			Full-context			NeSy		
Model	Co	Ex	Hy	Co	Ex	Hy	Co	Ex	Hy	Co	Ex	Hy
GPT-4o	62.5	40.0	25.0	57.5	16.0	30.0	62.5	43.8	32.5	**72.5**	46.9	30.0
GPT-5.2 Chat	65.0	26.9	9.1	70.0	25.0	27.0	70.0	**62.9**	34.2	**72.5**	59.5	**35.0**
Llama 4 Maverick	60.0	33.3	18.9	37.5	16.0	30.8	60.0	51.4	**37.5**	**72.5**	**55.3**	25.0
Claude 3.7 Sonnet	65.0	36.8	25.0	65.0	23.1	22.2	70.0	52.8	26.3	**70.0**	**53.8**	**30.0**
Claude 4 Sonnet	57.5	8.3	14.3	67.5	20.8	27.8	62.5	**56.8**	25.6	**70.0**	45.0	**27.5**
Claude 4 Opus	60.0	14.8	23.1	62.5	20.8	27.0	**70.0**	**57.9**	**30.0**	67.5	48.7	27.5
DeepSeek V3	45.0	40.0	25.0	57.5	28.6	26.3	55.0	48.6	27.5	**67.5**	**46.2**	**27.5**
DeepSeek R1	57.5	38.5	21.1	**62.5**	31.0	27.5	60.0	**60.0**	**32.5**	60.0	51.4	26.3
Qwen 2.5 72B	45.0	36.8	21.1	55.0	16.7	21.2	45.0	51.5	30.0	**67.5**	**57.1**	25.0
Gemini 2.5 Flash	60.0	41.0	25.0	57.5	20.0	28.9	62.5	**55.9**	**30.0**	**65.0**	52.8	25.0
Llama 3.1 8B	17.5	40.0	3.2	30.0	12.5	19.4	22.5	15.4	25.0	**65.0**	**48.7**	**25.0**
Llama 3.3 70B	45.0	42.5	21.1	52.5	16.0	19.4	57.5	50.0	27.5	**62.5**	**52.8**	**35.0**

Table 4: Complete 3-tier evaluation (%) across 12 LLMs and 4 strategies. Co = core genre; Ex = extension subcategory (24/40 acts have ground truth, percentage is calculated on this number); Hy = hybrid tags. Bold = best strategy per model per tier.

reason behind misclassification. The texts are full of verbs about giving, receiving, or appointing, as well as references to property, which can easily fool the model. The Toxotis documents present a consistent pattern of procedural misidentification. The models seem to get carried away by standard diplomatic wording, particularly the delegation formulas, and immediately assume that they are dealing with a transaction or a representation act. They completely miss the fact that the document is merely setting up someone's participation in a specific legal procedure. This inflates the representation of certain categories and under-identifies procedural acts such as Declaratio, Arbitrium, and Renuntiatio.

A related phenomenon is what may be termed "frame dominance," in which the diplomatic structure of an instrument outweighs its legal effect in the model's prediction. Despite frequently containing formulaic language associated with contracts, the texts from Toxotis are procedural in terms of the resulting act. This issue reveals a central problem for legal NLP in historical diplomatic materials. It shows that the surface wording of an instrument does not reliably reflect its true legal function.

The Toxotis register is also important from the standpoint of the NeSy system presented in this paper, i.e. it showcases that symbolic rules keyed to legal effect bypass the diplomatic framing that misleads the classifier by surface lexical cues and formulaic language. Procedural texts are a useful stress test for NeSy precisely because improvements there cannot be explained by pattern memorization alone.

Concluding, we can say that the Toxotis examples can function as a stress test for the design of our taxonomy, as well as annotation stability, since it reveals category boundaries that are more difficult to see in more standard transactional registers. Furthermore, these findings elucidate the overall performance trends, as they show that the

performance of a model might shift systematically depending on the diplomatic register and that a high density of procedural language heavily impacts category recognition. In the end, the Toxotis examples seem to back up our broader point: in order to reliably classify historical legal documents, one has to identify the primary legal effect instead of just relying on vocabulary or standard formulas.

7. Conclusion and Future Work

In this paper, we presented HeptaTAX, which consists of the following: a) a corpus of 1,088 sixteenth-century Corfiot notarial acts, b) a 40-act benchmark with 3-tier genre annotations, c) a taxonomy of 17 core legal genres, and d) a neuro-symbolic pipeline for automatic classification. The NeSy pipeline achieves 72.5% core accuracy, a 14.4% improvement over the mean zero-shot baseline, and compresses the accuracy gap between the strongest and weakest models from 47.5 pp to 12.5 pp. The symbolic engine acts as an equaliser across model sizes: Llama 3.1 8B gains 47.5 pp with NeSy and matches frontier models operating without symbolic support. To our knowledge, HeptaTAX is the first computationally accessible corpus of Heptanesian notarial acts with structured genre annotations. The NeSy approach is applicable to other historical document classification tasks where domain expertise can be formalised as rules, representing a form of knowledge transfer from humanities to NLP.

Beyond its computational contribution, HeptaTAX provides a structured framework for linking documentary typology with formulaic language, offering a reproducible basis for future linguistic and dialectological investigation of Early Modern Greek.

For future work, we will provide a full annotation of the whole corpus (1088 acts) in a semi-automated way, using the best-performing NeSy pipeline for pre-annotation followed by human correction. We will expand the symbolic rules to cover rare categories and procedural acts. Furthermore, we would like to extend this approach to other territories that were under Venetian rule and have similar notarial traditions (e.g. Crete, Cyprus. Lastly, we would like to attempt fine-tuning of a small model on the annotated data to see the effect of fine-tuning and how well this fares when compared to prompt engineering and our NeSy system.

Another relevant research direction is systematically modelling lexical bundles and diplomatic formulas in the form of signals that link surface textual patterns to juridical effect. The results presented in this paper point to formulaic language playing a central role in shaping classification behaviour (particularly in procedurally dense registers). The explicit representation of formula structures, whether statistically and/or symbolically, can provide an intermediate layer that is able to connect textual variation with legal function. Lastly, implementing juridical effect as a computationally tractable category allows one to connect historical diplomatics with NLP and shows that legal function (and not merely surface wording) can be effectively used in document classification.

In a broader sense, this type of research points to a research agenda where historical legal NLP goes beyond document classification and moves towards modelling the interaction between formulaic expressions, diplomatic register, and juridical effect. Thus, this approach can contribute to more robust effect-oriented classification and to legal NLP systems that are more explainable and transparent, given their grounding in historical documentary practice.

8. Limitations

The 40-act benchmark is small; a larger annotated set would strengthen the claims. The rules are hand-crafted and specific to the Venetian-Greek notarial tradition and would need adaptation for other legal traditions or periods. Temperature was set to 0.0 throughout, meaning each model produced a single deterministic run with no variance analysis. It would be interesting to check temperature variations across many different runs. The extension and hybrid classification tasks require more refinement, with their best scores being at $\sim$63% and $\sim$35%, respectively. Lastly, the Toxotis case shows that procedural registers require dedicated annotation and rule development that are lacking in the current work.

9. Bibliographical References

Antonios Anastasopoulos, Alessandro Cattelan, Kevin Duh, Christian Federmann, and Chris Callison-Burch. 2018. Part-of-speech tagging on an endangered language: A parallel Griko-Italian resource. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 29–35. Association for Computational Linguistics.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and curriculum-based writing. *Applied Linguistics*, 25(3):371–405.

Stavros Bompolas. 2023. *Computational Dialectology in the Linguistic Varieties of Cappadocian, Pharasiot, and Silliot*. Ph.D. thesis, University of Patras.

Stavros Bompolas, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2025. Crossing dialectal boundaries: Building a treebank for the dialect of Lesbos through knowledge transfer from Standard Modern Greek. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 39–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Stergios Chatzikyriakidis and Shalom Lappin. To appear. Neuro-symbolic NLP: Taxonomy, assessment, and directions. *Frontiers in Artificial Intelligence*. Research Topic on Neural-Symbolic NLP: Bridging Theory and Practice.

Stergios Chatzikyriakidis and Anastasia Natsina. 2026. LLMs got rhythm? Hybrid phonological filtering for Greek poetry rhyme detection and generation. In *Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026)*. Association for Computational Linguistics.

Stergios Chatzikyriakidis, Dimitris Papadakis, Sevasti-Ioanna Papaioannou, and Erofili Psaltaki. 2026a. GRDD+: An extended Greek dialectal dataset with cross-architecture fine-tuning evaluation. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026)*.

Stergios Chatzikyriakidis, Erofili Psaltaki, Dimitris Papadakis, Erik Henriksson, and Veronika Laippala. 2026b. Perplexity as a metric for dialectal distance: A computational study of Greek varieties. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026), EACL 2026*.

Stergios Chatzikyriakidis, Chatrine Qwaider, Ilias Kolokousis, Christina Koula, Dimitris Papadakis, and Efthymia Sakellariou. 2023. GRDD: A dataset for Greek dialectal NLP. In *Proceedings of the 16th International Conference on Greek Linguistics (ICGL16)*.

Luciana Duranti. 2015. *Diplomatics: New Uses for an Old Science*. Rowman and Littlefield, Lanham, MD.

Monireh Ebrahimi, Pascal Hitzler, Md Kamruzzaman Sarker, et al. 2024. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web*, 15(4):1291–1329.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2).

Aaron X. Fellmeth and Maurice Horwitz. 2021. *Guide to Latin in International Law*, 2 edition. Oxford University Press.

Silvia Gasparini. 2023. Notaries and the law in Venice: Development of a discipline. *Italian Review of Legal History*, 9:1–33.

Harris Hadjidas and Maria C Vollmer. 2015. Multi-CAST Cypriot Greek. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*.

Eleni Kakoulidi-Panou, Eleni Karantzola, and Katerina Tiktopoulou. 2023. *Vernacular Prose of the 16th Century*. Greek Language Center, Athens. In Greek.

Eleni Karantzola. 2024. What do we know about and what can we learn from Early Modern Greek? In *Proceedings of the 15th International Conference on Greek Linguistics*, volume I, pages 49–68, Belgrade. University of Belgrade. In Greek.

Eleni Karantzola and Nikolaos Lavidas. 2016. Characteristics of 16th-century Corfiot. *Studies on the Greek Language*, 36:129–150. In Greek.

Eleni Karantzola and Vasiliki Makri. in print. A comparative analysis of 16th-century Corfiot and Cephalonian: Evidence from notarial records. In *Proceedings of Modern Greek Dialects and Linguistic Theory 10*.

Eleni Karantzola, Katerina Tiktopoulou, and Katerina Frantzi. 2012. Notarial documents as sources for Early Modern Greek. In *Early Modern Greek Public Secretariat Language, Tradition, and Poetics*, pages 473–501. Vikela Municipal Library, Heraklion. In Greek.

Nikos Liosis. 2024. Heptanesian dialects. In G. Giannakis, editor, *Encyclopedia of Greek Language and Linguistics Online*. Brill.

Vasiliki Makri. 2020. *The Formal Expression of Grammatical Gender in a Modern Greek Dialect Affected by Italo-Romance*. Ph.D. thesis, University of Patras, Department of Philology.

Io Manolessou. 2003. Non-literary sources as evidence for the language of the medieval period. *Lexicographic Bulletin*, 24:61–88. In Greek.

Erofili Psaltaki, Yves Scherrer, and Stergios Chatzikyriakidis. 2025. Italian and Turkish loanwords detection in Greek dialects. In *Proceedings of the 17th International Conference of Greek Linguistics (ICGL 2025)*, Cambridge, UK.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint*

25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 776–784, Tallinn, Estonia. University of Tartu Library.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech recognition for Greek dialects: A challenging benchmark. In *Interspeech 2024*, pages 3974–3978.

Esther-Miriam Wagner, Ben Outhwaite, and Bettina Beinhoff, editors. 2013. *Scribes as Agents of Language Change*. Mouton de Gruyter, Berlin.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge.

Towards Semantic Access and Interoperability in Digital Dialectal Atlases. A Case Study

Paola Marongiu, Simonetta Montemagni
Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche (CNR-ILC)
Via Giuseppe Moruzzi 1, Pisa
{paola.marongiu, simonetta.montemagni}@ilc.cnr.it

Abstract

The increasing digital availability of dialectal atlases has significantly enhanced access to dialectal data and their potential for linguistic and cultural studies. However, despite their richness, such resources often remain difficult to integrate into contemporary data-driven research workflows, due to complex data structures and limited interoperability. Most digital dialectal atlases still rely on traditional access models centered on maps, offering only implicit and coarse-grained semantic structures, which limits concept-based exploration and potential for integration with other linguistic resources in the Linguistic Linked Open Data (LLOD) ecosystem. This paper presents a case study carried out on the *Atlante Lessicale Toscano* (ALT), aimed at addressing these limitations through the introduction of an explicit semantic layer designed to support both user-oriented exploration and machine-actionable interoperability. While ALT already provides a conceptual organization of dialectal materials, this structure was originally conceived for human navigation and not for integration with other computational lexical-semantic resources. To bridge this gap, we align ALT concepts with ItalWordNet, leveraging its synset-based model as a widely adopted semantic backbone in NLP and LLOD infrastructures. The case study focuses on the domain of agriculture, whose historically grounded conceptual distinctions are often underrepresented in general-purpose lexical resources. The paper proposes a mapping strategy, analyzes coverage and mismatch patterns, and releases a new aligned resource mapping ALT agricultural concepts to ItalWordNet, thereby creating the prerequisites for interoperability and reusability of dialectal atlas data.

Keywords: dialectal resources, WordNet, computational lexical semantics, specialized knowledge

1. Introduction

In recent years, a growing number of dialectal atlases have been made available in digital form, substantially enhancing access to and exploration of dialectal data. These resources play a key role in the preservation of cultural heritage and domain-specific knowledge. Beyond their linguistic relevance, dialectal atlases encode fine-grained conceptual distinctions related to traditional practices, environmental knowledge, and material culture, thereby representing a valuable source of cultural information. However, such resources often remain difficult to integrate into contemporary data-driven research workflows and limited interoperability.

We focus here on resources documenting dialects of Italy: notable examples include two digital versions of the *Sprach- und Sachatlas Italiens und der Südschweiz* (Atlante Italo-Svizzero, AIS, 1928-1940), namely *NavigAIS*[1] (Tisato, 2019) and *AIS Reloaded*[2] (Loporcaro et al., 2021), as well as atlases circumscribed to specific areas, such as the *Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects, ALD*[3] (Goebl, 2020) or the *Atlante Lessicale Toscano* (ALT)[4] (Picchi et al., 2001; Cucurullo et al., 2006). All of them are currently accessible online.[5] While their digitalization significantly enhances accessibility from a human-user perspective, it does not automatically guarantee compliance with FAIR principles (Wilkinson et al., 2016) or readiness for integration within the Linguistic Linked Open Data (LLOD) ecosystem.

Despite the growing availability of digital dialectal atlases, access to dialectal data largely mirrors the intrinsic organization of traditional linguistic atlases. To the best of our knowledge, with the partial exception of ALT (see below), none of the above mentioned resources currently provides a structured and machine-actionable semantic layer enabling concept-based queries. The primary access key typically remains the map, retrieved via a numerical identifier or title. As a consequence, dialectal data are findable and accessible in a broad sense, but only weakly interoperable and reusable in the sense defined by the FAIR principles. In particular, the absence of shared ontologies, standardized vocabularies, and Linked Data representations limits

[1] https://navigais-web.pd.istc.cnr.it/

[2] https://www.ais-reloaded.uzh.ch/

[3] https://www.ald.gwi.uni-muenchen.de/it/?db=ald1

[4] https://dbtvm1.ilc.cnr.it/altweb/RT_ALT-WEB_home.htm

[5] More recent initiatives include the digitization of the *Atlante Linguistico Italiano, ALI*, a foundational resource for dialectological research on Italian dialects which, however, is still ongoing (Cerruti et al., 2025).

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 274–286
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

their integration with external datasets and reduces their potential for computational reuse.

Semantic organization in dialectal atlases is generally limited to broad semantic fields, which often correspond to individual atlas volumes. For instance, the AIS questionnaire was organized into eight parts, each including approximately 200 questions (therefore maps). While this coarse-grained and essentially flat organization may offer initial guidance within the atlas, it is not explicitly represented or operationalized in its digital counterparts. Moreover, such groupings tend to be highly heterogeneous: for example, AIS Volume 1 encompasses topics ranging from relationships, age, and love to birth, marriage, death, given names, body parts, body functions, and physical qualities and defects. This, in turn, hinders concept-driven exploration of the atlas and complicates its mapping to external semantic resources.

The picture sketched above highlights that the digital availability of dialectal atlases raises new challenges concerning semantic organization, conceptual accessibility, and interoperability, which represent core requirements for FAIR-compliant data. Addressing these challenges serves a twofold purpose: internally, it supports users in navigating dialectal data along explicit semantic paths; externally, it establishes the prerequisites for integration with other linguistic resources and tools. For dialectal resources to remain relevant and usable within this evolving ecosystem, they must become interoperable with computational infrastructures and widely adopted encoding and semantic standards. Against this background, this paper presents a case study carried out on the ALT dialectal resource, aimed at addressing these limitations through the introduction of an explicit semantic layer, designed both to enhance user-oriented exploration and to support machine-actionable interoperability.

It is worth noting that ALT already provides a conceptual organization of dialectal materials, articulated into macro-classes, micro-classes, and concepts corresponding to individual questions of the ALT questionnaire (Cucurullo et al., 2006). However, this organization was conceived primarily to facilitate access and navigation for human users and was not designed for integration with other computational resources. To bridge this gap, we decided to use WordNet (WN) — specifically ItalWordNet (IWN) (Roventini et al., 2016; Monachini et al., 2016). This choice is motivated by the widespread adoption of WN as a lexical-semantic backbone in NLP and LLOD infrastructures (Litta et al., 2025; Mambrini et al., 2021), its synset-based model enabling interoperable semantic mapping, and its suitability as a reference framework to assess the alignment between domain-specific dialectal concepts and general-purpose lexical knowledge. For our case study, we decided to focus on the domain of agriculture, whose meanings and conceptual distinctions reflect historically grounded categorizations that are often underrepresented or absent in contemporary general-purpose lexical resources.

Our contributions are threefold: (i) we define a strategy to tackle the methodological challenges involved in mapping domain-specific concepts to a general-purpose lexical-semantic resource; (ii) we provide a quantitative and qualitative analysis of concept coverage and mismatch patterns, highlighting gaps and differences in granularity between dialectal and computational lexical-semantic resources; and last but not least (iii) we release a new aligned resource linking ALT agricultural concepts to IWN synsets, thereby enhancing the interoperability of dialectal data within the LLOD ecosystem.

2. The *Atlante Lessicale Toscano*

The ALT is a specialized linguistic atlas aimed at documenting lexical variation across both diatopic and diastratic dimensions. Data were collected between 1974 and 1986 through a large-scale fieldwork campaign carried out by trained interviewers in 224 locations throughout Tuscany, a region of central Italy. A total of 2,193 informants were selected according to standard sociolinguistic criteria, including age, educational level, socio-economic status, and cultural background.

2.1. The Questionnaire

Data elicitation relied on a questionnaire comprising 745 items, primarily targeting variation in lexical choice, meaning, and pronunciation. Two main elicitation strategies were adopted. On the one hand, onomasiological questions started from a given concept and asked informants to provide the corresponding lexicalization, e.g. terms used to denote 'bread crumbs'. On the other hand, semasiological questions began with a dialectal form and required informants to specify its meaning or range of referents together with its pronunciation, e.g. the term *ceppo*, which in Tuscan varieties may denote a 'tree stump', a 'log', or a 'Christmas present'.

Answers to both types of questions were phonetically transcribed and enriched with informants' explanations, comments, and any additional linguistic material that emerged during the interviews, even when not directly related to a specific questionnaire item. Overall, the data collection process yielded several million individual responses, subsequently organized into approximately 380,000 structured records. Of these, about 350,000 correspond to canonical responses to questionnaire items attested in the surveyed locations - often accompanied by usage contexts and metalinguistic

remarks - while roughly 30,000 records document dialectal forms collected as supplementary material during fieldwork.

2.2. ALT Versions

ALT was first released in 2000 as a CD-ROM (Giacomelli et al., 2000), providing users with the possibility to explore dialectal data through complex, parameter-driven queries tailored to specific research needs. With the increasing availability of internet-based technologies, the original CD-ROM edition was eventually superseded by ALT-Web[6] (Montemagni et al., 2006), an online platform granting access to the full body of linguistic data collected for the atlas. This transition substantially expanded the potential user base, reaching not only dialectologists and linguists but also non-academic users interested in the linguistic and cultural heritage of Tuscany.

In the CD-ROM edition of the ALT, an attempt was made to address the difficulty of retrieving specific concepts from the questionnaire through a set of 338 keywords designed to identify thematic groupings of questions. Although this provided some support for navigating the questionnaire, it still required consulting the entire list. In ALT-Web, by contrast, an ontology-based search was introduced, representing a significant enhancement that substantially improved user accessibility.

ALT-Web should still be regarded as a legacy application, together with its associated data. A new version of the resource, named ALT-Web OPEN, is currently under development with the aim of complying with the FAIR principles (Cucurullo et al., 2025). The overarching objective of this initiative is to ensure the long-term preservation of ALT dialectal data, to improve sustainable and open access, and to enable interoperability with other linguistic resources and computational tools. The work presented in this paper is part of this process.

2.3. ALT-Web Conceptual Organization

One of the canonical access keys to the materials of a linguistic atlas collected through a questionnaire is question-based access: within the corpus, all attestations elicited in response to a selected question are retrieved. This type of access, however, presupposes familiarity with the questionnaire used for data collection, which is often articulated into hundreds of questions. As mentioned above, in the case of ALT, the data collection questionnaire includes 745 items.

[6] https://dbtvm1.ilc.cnr.it/altweb/RT_ALT-WEB_home.htm

2.3.1. ALT-Web Ontology

To further support users of ALT-Web in identifying the question or set of questions relevant to their interests, an ontology was developed to organize the concepts investigated by ALT into a domain-based hierarchy. In particular, the set of concepts covered by the ALT questionnaire was organized into 13 macro-classes, derived from the original subdivision of the questionnaire into thematic sections, plus an additional miscellaneous class collecting questions not readily assignable to the identified macro-classes.

At the highest level of the ALT ontology are the macro-classes listed in the first column of Table 1. For each macro-class, a set of intermediate conceptual classes — hereafter referred to as micro-classes corresponding to more fine-grained semantic groupings — has been identified (reported in the second column of the table), yielding a total of 396 such associations. On average, each macro-class is articulated into 28 micro-classes. At a lower level of the ontology, each micro-class is associated with a set of low-level keywords (inherited from the CD-ROM version), which serve to define more specific semantic groupings within a micro-class. These are, in turn, linked either to (a) elementary concepts expressed by Italian lexical items, including both single words and multi-word expressions, corresponding to onomasiological questions (third column), or to (b) dialectal lexical items, corresponding to semasiological questions (fourth column). The terminal nodes of the ontology thus correspond to the questionnaire items themselves. The final column of Table 1 reports the number of items associated with each macro-class, for a total of 2,750 associations.

2.3.2. Organization Criteria

The criteria used to associate ALT questionnaire items with the conceptual classes of the ontology vary according to the type of question. Onomasiological questions are linked to the elementary concepts they investigate, which are in turn associated with progressively broader conceptual groupings up to the top-level nodes of the ontology. For onomasiological questions, each questionnaire item is classified under a single macro-class, while different semantic facets or possible cases of polysemy are captured through the association of the same question with multiple micro-classes. For example, question 169, aimed at collecting the lexicalizations for the concept 'head pad', is recorded in two different micro-classes under the macro-class AGRICULTURE, that of TOOL and that of TRANSPORTATION, highlighting different meaning facets (corresponding to "hyperonymy" and "used_for" relations).

In the case of semasiological questions, the pat-

Macro-class	Associated micro-classes	Onom. quest. items	Semas. quest. items	All quest. items
AGRICULTURE	39	124	220	344
FOOD AND NUTRITION	31	151	148	299
ANIMAL HUSBANDRY	16	67	87	154
WILD FAUNA	11	49	44	93
FORESTRY AND WOOD HARVESTING	23	60	145	205
HOUSING AND DOMESTIC ACTIVITIES	36	109	147	256
LANDFORMS AND GEOMORPHOLOGY	26	57	78	135
PLANTS AND FRUITS	23	115	60	175
CHRONOLOGICAL TIME	8	21	18	39
WEATHER AND METEOROLOGICAL PHENOMENA	17	64	35	99
HUMAN ACTIVITIES AND SOCIAL RELATIONS	32	68	90	158
HUMAN BEHAVIOUR AND EMOTIONS	70	84	316	400
HUMAN BODY AND CLOTHING	55	130	238	368
MISCELLANEOUS	9	7	18	25
Total	396	1106	1644	2750

Table 1: ALT-Web ontology: macro-classes and associated micro-classes and questionnaire items (onomasiological vs semasiological).

tern of associations at the terminal and pre-terminal levels of the ontology is completely different. Semasiological questions are connected to broader conceptual classes via the dialectal term under investigation. These associations were established based on the analysis of the meanings attested for each dialectal term in Tuscany. Therefore, a single semasiological question may be classified as relevant to multiple macro-classes. For instance, question 434d, aimed at eliciting the meanings of the dialectal term *manfano*, is associated — due to the marked polysemy of the term in Tuscan dialects — with the macro-classes AGRICULTURE, FORESTRY AND WOOD HARVESTING, PLANTS AND FRUITS, HUMAN BEHAVIOUR AND EMOTIONS, and HUMAN BODY AND CLOTHING. This distinction within the ALT ontology is shown in Figure 1 with reference to the semasiological question 434d *manfano*, whose attested meanings range across different macro-classes.

2.3.3. Ontology-based Search

This organization of the questionnaire enables concept-based access to dialectal materials. Users are first presented with the list of conceptual macro-classes; once a relevant macro-class has been selected, a set of micro-classes corresponding to more fine-grained conceptual groupings is displayed. By selecting a specific concept, users can then access the set of associated questionnaire items, further subdivided into two subsets according to question type. Figure 2 illustrates the results of a query targeting the micro-class TOOL, as displayed by the ALT-Web interface, in which questionnaire items are organized into two groups according to question type.

Unlike a flat list of lexical items, which merely provides keywords for retrieval, a conceptually structured resource offers an explicit network of relations among items, enabling more informed semantic interpretation. The ALT-Web ontology, therefore, functions as an effective, access-oriented semantic layer, designed to facilitate navigation and retrieval rather than to provide a logically complete or formally constrained conceptual model.

Notably, the same conceptual entity may appear under multiple macro-classes, reflecting alternative interpretative perspectives rather than a strict subsumption-based hierarchy. While this design choice enhances usability and supports exploratory access, it limits the degree of ontological systematicity that can be achieved, and has direct implications for the integration of external semantic resources. Our mapping onto IWN is intended to address this issue.

3. WordNet

The choice of WordNet (WN) as the target semantic resource for our mapping is not self-evident. The most appropriate strategy for linking dialectal atlas data with computational semantic resources remains an open issue. Existing approaches include the alignment of entries with Wikidata entities, as adopted in *VerbaAlpina* (Colcuc, 2020), or with Concepticon identifiers (List et al., 2016), as in the digitization of the *Tableaux Phonétiques* (Hans Geisler, 2021). In both cases, the linking strategies rely on flat inventories of entities or concepts associated with shared identifiers, rather than on an explicitly articulated conceptual hierarchy.

While such approaches effectively establish cross-resource correspondences, they do not provide a formal representation of conceptual organization or of the semantic relations connecting lexical meanings. For modeling dialectal atlas data, where lexical variation frequently reflects subtle semantic distinctions and patterns of polysemy, a

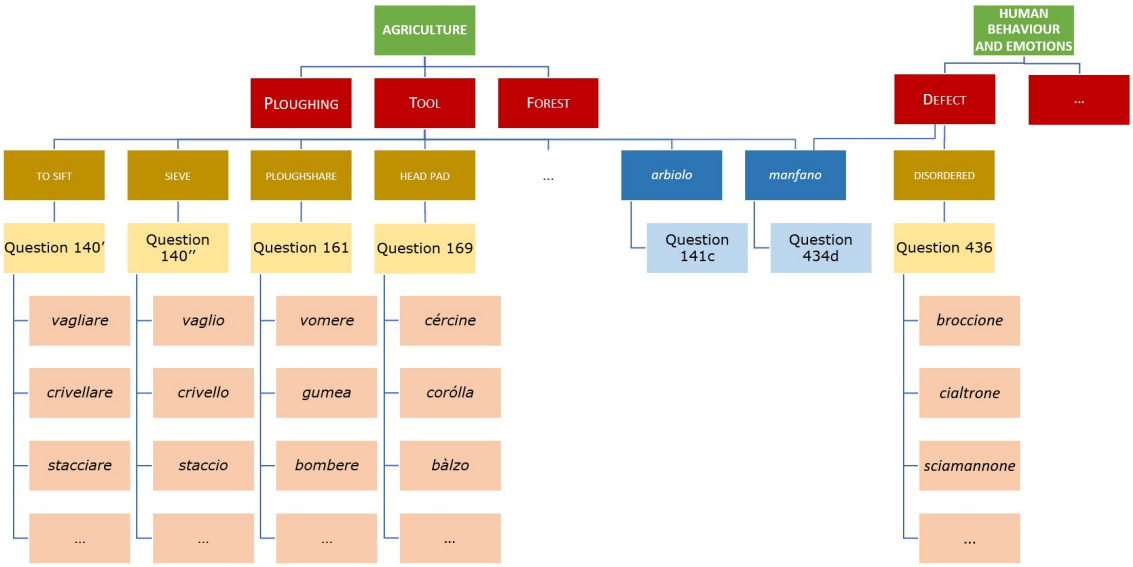


Figure 1: Fragment of ALT-Web ontology with onomasiological and semasiological questions, marked respectively by yellow and light blue boxes.

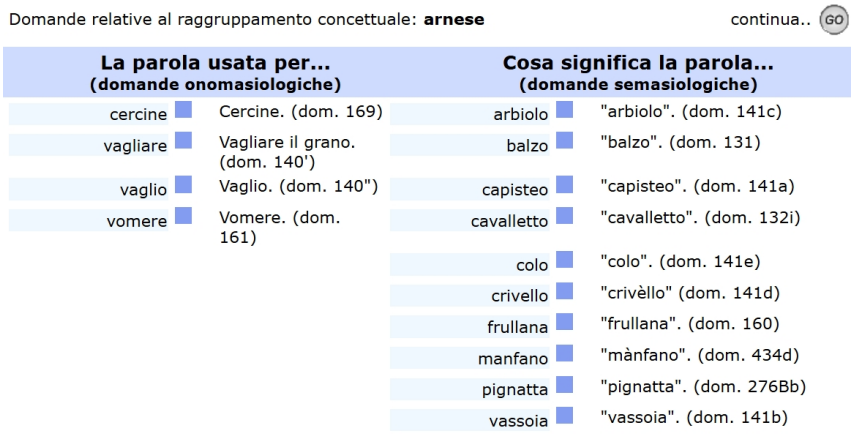


Figure 2: ALT-Web query results for the micro-class TOOL: onomasiological (left) and semasiological (right) questions. The keyword associated with each question is shown on the left: Question 140" "Vagliare il grano" 'to winnow grain' is associated to the keyword "vagliare" 'to winnow'.

structured lexical-semantic architecture appears more suitable. Moreover, in accordance with the FAIR principles, semantic interoperability requires the use of formal, shared, and machine-readable knowledge representations. A relational conceptual model provides a stronger foundation for both interoperability and reusability than simple identifier alignment.

Within this perspective, WN constitutes a particularly relevant modelling framework. Princeton WordNet (Fellbaum, 1998; Miller et al., 1990) was conceived as a lexical-semantic network in which meanings are represented not as isolated dictionary entries, but as nodes in an interconnected conceptual system. The fundamental unit of this architecture is the *synset* (synonym set), which corresponds to a single lexicalized concept. Each synset groups together lexical items (lemmas) that are interchangeable in a given context and thus express the same sense. WN draws a principled distinction between lemmas and senses: while a synset encodes one sense, a lemma may belong to multiple synsets. Polysemy is therefore modelled as the distribution of lemmas across distinct conceptual nodes, rather than as multiple definitions attached to a single lexical entry. For example, the English word *bank* appears in different synsets corresponding to a financial institution and to a river bank (Miller et al., 1990). Each synset is identified by a unique offset and is accompanied by (i) a gloss provid-

ing a definition of the concept, (ii) usage examples when available, and (iii) a set of formally typed semantic relations connecting it to other synsets. The most central relation is hypernymy/hyponymy, which structures the network hierarchically. Other relations include meronymy/holonymy (part–whole relations), as well as similarity and antonymy relations. This relational layer transforms WN into a structured semantic graph in which concepts are embedded within a formally defined hierarchy.

On the basis of the Princeton model, several language-specific WNs have been developed. Among them, Italian WordNet (IWN) (Roventini et al., 2016; Monachini et al., 2016) is a manually constructed lexical-semantic database for Italian, built in alignment with the original architecture while accounting for language-specific lexical and semantic distinctions. For instance, in IWN the synset with offset `#12355-n`, glossed as "mobile costituito da un piano sostenuto per lo più da quattro gambe" 'piece of furniture consisting of a flat surface typically supported by four legs', is associated with the lemma *tavolo* 'table' and is linked via hyperonymy to `#20176-n` "oggetto d'arredamento" 'furniture item', which in turn connects to more abstract nodes such as `#4716-n` "oggetto materiale, concreto" 'material, concrete object' and ultimately `#15544-n` "ciò che esiste" 'existing entity'.

Aligning dialectal concepts to WN is not intended to reduce their domain-specific richness. Rather, it anchors them within a shared and formally structured semantic space, thereby facilitating FAIR-oriented interoperability, cross-resource querying, and computational reuse.

4. Mapping ALT to IWN

Each level of the ALT ontology for the agricultural domain was mapped to a synset in IWN. The mapping process was carried out entirely manually, and involved all levels of the ALT-Web ontology for the agricultural domain. We started from the macro-class AGRICULTURE; then we mapped the micro-classes belonging under AGRICULTURE (39), e.g. PLOUGHING, TOOL, FOREST, etc.; finally, we mapped the concepts corresponding to onomasiological questions for each micro-class e.g., for the micro-class TOOL, VAGLIARE 'to winnow', VAGLIO 'sieve', VOMERE 'ploughshare', CERCINE 'head pad' etc. At this stage, semasiological questions have not been included in the mapping process, as the present study primarily aims to align the concepts investigated through the ALT questionnaire with IWN. The treatment of semasiological questions — namely, the association between dialectal items and their meanings as attested in Tuscany, already codified within the ALT ontology — will be deferred to a subsequent phase, contingent upon an assessment of the feasibility of mapping ALT concepts to IWN. The onomasiological questions for the macro-class AGRICULTURE are shown in Table A1 in Appendix.

In mapping a concept to a synset, we rely on the full text of the question that explicates the concept, rather than on the keyword alone (refer back to Figure 2). For instance, the keyword *grappolo* 'bunch' is associated with three distinct questions, and therefore with three different concepts: Question 115 GRAPPOLO D'UVA 'bunch of grapes'; Question 117 PARTE DEL GRAPPOLO D'UVA 'part of a bunch of grapes'; and Question 118 DOPPIO GRAPPOLO 'double bunch'. Each question is thus linked to a distinct synset, even when the keyword is identical. In order to identify the appropriate synset for mapping the concepts of the ALT ontology, we searched for corresponding concepts in version 2.0 of IWN (Monachini et al., 2016), querying both lemmas and glosses of the synsets.

The mapping process for such domain- and culturally specific concepts raised some challenges which required us to devise strategies to tackle them. These challenges can be grouped into different categories: i) extremely specific concepts that appear in ALT but do not have a corresponding synset in IWN (section 4.1) ii) multiple synsets available for the same concept (section 4.2) iii) ambiguous concepts that might be relevant to more than one synset (4.3).

4.1. Missing synsets for domain-specific concepts

In some cases, the dialectal material describes extremely fine-grained domain-specific concepts that do not have a directly corresponding synset available in IWN. This is the case for the concept VAGLIARE IL GRANO 'to winnow grain' (Question 140'), which is lexicalized in various ways in Tuscan dialects, e.g. *burattàre, conciàr l grào, nettàre l gràne*. In Italian however, a specific lexeme to express this specific action is missing, and this gap is reflected by IWN. A blatant example of fine-grained concept distinctions is given by questions 115, 117 and 118, referring, respectively, to bunch of grapes, part of a bunch of grapes, and double bunch of grapes (see above), all three lexicalized in different ways across Tuscany. In order to keep track of such problematic cases, we opted for a two-level mapping: in the first level, we give the value "NA" to express the absence of a specific synset for that concept; in the second level, we map the concept onto the closest synset available in the IWN hierarchy. For instance, we map *vagliare il grano* on the second level onto a more generic synset `#38290-v` "vagliare" 'to examine'. Table 2 shows the two-level mapping for the concepts associated with the keyword *grappolo*.

It is worth mentioning that in some cases for an extremely specific concept an equally specific synset was found. That was the case for the concept RACCOGLIERE, which in ALT is paired with lexical entries describing the action of harvesting, inherently specific to agricultural activities. In this case, an extremely specific synset was found in IWN, paired with the lexical entry *raccogliere*, glossed as "prendere e radunare i frutti della terra" 'harvesting and gathering the fruits of the land'.

4.2. Multiple synsets available

In contrast, in some other cases, more than one synset could be mapped onto the same ALT concept. For instance, GRANO 'wheat' was relevant to three different synsets: `#26302-n` "pianta erbacea con foglie lineari e infiorescenza a spiga" 'herbaceous plant with linear leaves and a spike inflorescence'; `#26303-n` "frutto del grano da cui si ricavano farine alimentari; frutti cariossidi ('chicchi') da cui si ricava farina per pane e paste alimentari" 'fruit of wheat from which food flours are obtained; caryopsis fruits ('grains') from which flour for bread and pasta is produced'; `#8172-n` "pianta che produce cariossidi gialle" 'plant that produces yellow caryopses (grains)'. In such cases, we decided to map GRANO onto a primary synset, and to provide a second and third option of mapping to preserve coherence and ensure the possibility of exploiting the IWN hierarchy in all three cases. It should be noted that *grano* and *granoturco* (or *granturco*) 'corn' in IWN belong to the same synset (`#8172-n`), despite technically referring to two different types of grains (wheat vs. corn). In this case, we decided to preserve the IWN information as it is.

4.3. Conceptual collapse and polysemy in mapping

During the alignment process, we observed the presence of ambiguous concepts in the ALT hierarchy. A representative example are the questions investigating the concepts FRANTOIO and SECCATOIO DI CASTAGNE. In ALT, the concept FRANTOIO may refer either to the tool used to crush olives ('olive-press') or to the building where olive oil production takes place ('olive oil mill'). IWN distinguishes these meanings: the synset corresponding to *frantoio* refers to the machine used for crushing olives, while the building is described by a different synset (`#31669-n`), associated with the lemma *oleificio*. The ALT informants do not consistently separate these senses; rather, they reflect a conceptual blending of tool and location. For the time being, and in the absence of systematically disambiguated response forms, we mapped FRANTOIO to the synset describing the olive-oil ('macchina per la frangitura delle olive'), acknowledging that this choice only partially captures the semantic range observed in the data.

The case of SECCATOIO DI CASTAGNE presents a different challenge. Chestnut drying is an agricultural activity particularly common in the Tuscan-Emilian Apennine area. By inspecting the answers collected in ALT under the concept SECCATOIO, we found that, just as for FRANTOIO, the interviewees refer to a physical building (a dedicated construction or small house for chestnut drying), or to a specific tool or component used in the drying process (e.g., a stone disc or wooden apparatus). However, no direct synsets exist in IWN for any of these specific concepts. Several candidate hypernyms were considered for the second-level mapping described in section 4.1, in line with the internal structure of the ALT ontology: (i) a type of building (micro-class LUOGO 'place'), (ii) an instance of a drying process (ESSICCAZIONE), or (iii) an activity within chestnut cultivation (CASTAGNICOLTURA 'chestnut farming'). All three perspectives are semantically defensible. However, given the heterogeneity of the atlas responses, we opted to map the concept to the synset corresponding to *essiccazione* 'drying process', which captures the functional purpose common to all referents for this concept.

5. Results and analysis

We carried out an analysis of the results of our mapping on two axes. First, we measured to what extent the knowledge (i.e., the concepts) represented in ALT for the macro-class AGRICULTURE can be mapped onto a general-purpose computational semantic resource like WN (section 5.1). Then, we compared the knowledge organization in the ALT ontology (Figure 1) with the one extracted from IWN by leveraging the synsets relations (section 5.2).

5.1. Concept coverage

As described in Section 4, the mapping process raised different challenges due to the presence of very specific concepts in the ALT ontology. In order to understand to what extent a culture-specific resource like ALT can be mapped to a general-purpose resource like IWN, we measured the concept coverage by counting the number of concepts and micro-classes that we could be mapped to an IWN synset. The results are shown in Table 3. In order to give an account of the concept coverage, we decided to reflect the two-level mapping strategy explained in section 4.1 in our analyses. Out of 34 micro-classes, only for two of them we had to rely on a hypernym, using the second-level mapping. That was the case for CASTAGNICOLTURA 'chestnut farming' and PORZIONE DI FRUTTO 'fruit portion'. Concept mapping was more problematic:

Concept	Translation	Synset	Closest synset
GRAPPOLO D'UVA	bunch of grapes	#26969-n	–
PARTE DEL GRAPPOLO D'UVA	part of a bunch of grapes	NA	#26969-n
DOPPIO GRAPPOLO	double bunch (of grapes)	NA	#26969-n

Table 2: Two-level mapping for domain-specific concepts related to GRAPPOLO D'UVA 'bunch of grapes'.

some of them are very specific to Tuscan agricultural activities and artifacts, as described in section 4.1. Out of 38 concepts (i.e. questions), only 25 of them were mapped directly to their corresponding synset. The remaining 13 concepts were mapped on the second level to the closest synset found in IWN. Overall, 0.3% of the concepts and only 0.05% of the micro-classes did not find an exact correspondence with a synset. However, by using the two-level mapping, all concepts and micro-classes were successfully mapped to IWN.

5.2. Dialectal materials in the WordNet hierarchy

To explore the distribution of the ALT concepts in the IWN network, we went up the hypernyms chain defined within the WordNet structure, until we found the root synset for each mapping. Given our focus on the dialectal materials, we only worked on mapping of a synset to an ALT concept (in yellow in Figure 1), discarding micro-classes. In this section, we will i) give an overview of the concept specificity in the IWN network for all the concepts in the macro-class AGRICULTURE; ii) show an example of how the IWN hierarchy can support and/or clash with the handcrafted ALT ontology.

Concept specificity in the WN hierarchy. To determine whether the supposed cultural specificity of the concepts represented in ALT corresponds to a specificity of senses as they are represented in the WN network, we calculated the length of the hypernym chain for each mapping i.e., how many steps are required to go from the mapped synset to its root synset, that is, the most general one. The reasoning is that a synset positioned deep within the hierarchy (and therefore requiring a longer hypernym chain to reach the root synset) should correspond to a greater specificity of the concept described by the synset itself.

First, we only considered direct mappings i.e., concepts for which a synset was already available in IWN. The depth of each synset in the WordNet hierarchy per number of concepts associated is shown in Figure 3. Then, we included in the computation both first and second-level mappings i.e., concepts for which we had to rely on the closest available synset in the WN hierarchy due to the absence of a synset specific enough to describe the concept. The results are shown in Figure 4. For each second-level mapping we added an additional layer to make up for the missing synset.

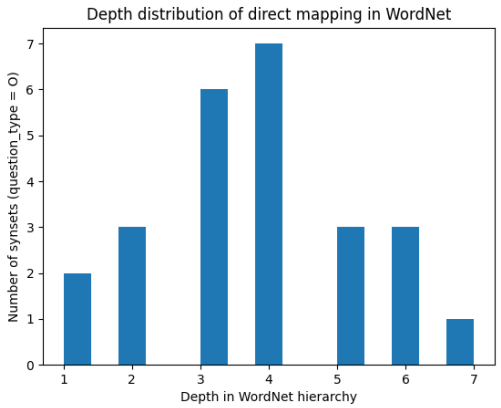


Figure 3: Synset depths for first-level mapping of concepts from ALT-Web for the macro-class AGRICULTURE.

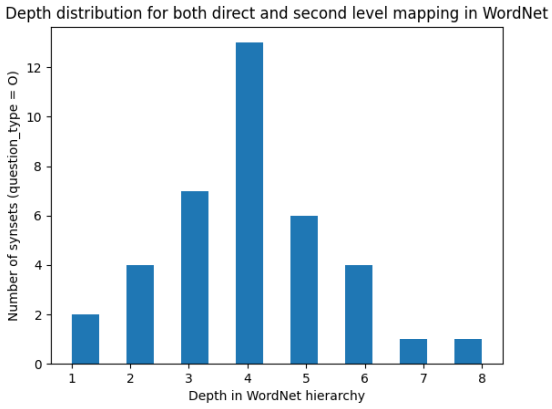


Figure 4: Synset depths for first and second-level mapping of concepts from ALT-Web for the macro-class AGRICULTURE.

In both cases, we find that the lexical layer of the ALT-Web ontology i.e., dialectal realizations in ALT-Web tends to align with more fine-grained WordNet nodes than the ontological key layer. As shown in Figures 3 and 4, most of the concepts (13 for first-level mapping only, and 20 for mapping at both levels) are located at the 3rd or 4th node of the semantic hierarchy in WordNet. A minority of concepts (7 for first-level mapping, and 10 for both levels) are located at the 5th, 6th, 7th, or event 8th node of the WordNet semantic hierarchy, suggest-

Level	Tot	Direct	Closest	% Closest
Micro-classes	34	32	2	0.05
Concepts	38	25	13	0.3

Table 3: Concept coverage for micro-classes and concepts in the agriculture macro-class (onomasiological questions only). Direct = concepts that were mapped directly to one synset; Closest = concepts that were mapped to the closest available synset.

ing that they refer to even more specific referents in the macro-class AGRICULTURE.

Concepts that appear deep in the WordNet hierarchy are related to words that describe specific activities in the field e.g. VAGLIARE IL GRANO 'to winnow grain' (offset `#38290-v`, 8th level of the hierarchy); MIETERE 'to reap' (offset `#32641-v`, level 6); TREBBIARE 'to thresh' (offset `#41163-v`, level 7). Concepts that appear at the 3rd and 4th level usually indicate similar activities, e.g. RACCOGLIERE LE OLIVE 'to harvest olives', probably less specific in the WN structure, or specific referents in the countryside: FASCIO DI SPIGHE LEGATE SUL CAMPO 'sheaf of ears of grain', PAGLIUZZA 'straw (small pieces)', RICCIO DELLA CASTAGNA 'chestnut burr'.

Handcrafted ontology vs. WordNet. We compared convergence and differentiation in IWN among the four concepts under the micro-class TOOL, shown in Figure 1: VAGLIARE IL GRANO 'to winnow grain', VOMERE 'ploughshare', VAGLIO 'sieve', CERCINE 'head pad'. The concepts VOMERE and VAGLIO display partially overlapping hypernymic trajectories, sharing higher-level synsets that correspond to general artifact and tool categories. This indicates that, despite referring to distinct implements, both items are semantically anchored in the same macro-domain of agricultural instruments. Their paths converge at abstract nodes encoding physical object and man-made artifact concepts, revealing a clear structural proximity in the WN hierarchy. By contrast, VAGLIARE IL GRANO, projects onto a different branch of the hierarchy, associated with action or process synsets rather than concrete artifacts. Its (second-level) mapping situates it within a procedural semantic field, thereby separating it structurally from the nominal tool-denoting items. This divergence reflects the ontological distinction between entities and events that is overlook in ALT. CERCINE follows yet another path. Although it ultimately converges with the other concepts at very abstract upper-level synsets (e.g., physical entity or artifact, depending on its mapping), its intermediate hypernyms differ, reflecting a more specialized or less centrally connected semantic niche within the agricultural ontology. Overall, we found that concepts belonging to the same thematic domain (agriculture) do not necessarily cluster at lower hierarchical levels. Instead, semantic convergence tends to emerge only at higher, more abstract synsets, while lower-level organization preserves fine-grained conceptual distinctions between tools, processes, and specialized objects.

It should be noted that, while WN provides a structured conceptual organization, its hierarchy does not necessarily reflect all possible perspectives on conceptual relatedness. Consider the concept ACINO 'grape'. IWN has a lexical entry for this word, assigned to two different synsets. One of the synsets is `#25757-n` "chicco dell'uva o di frutto simile" 'grape, or grain-like part of some other fruit'. Let us consider now the concept GRAPPOLO. This lemma has a lexical entry in IWN with which a few synsets are associated, again due to the polysemy of this word in Italian. The synset `#7317-n` "infruttescenza" 'infructescence' has the following hypernyms chain: infruttescenza > insieme dei frutti derivati da un'infiorescenza raggruppati in modo da sembrare un frutto unico 'a group of fruits derived from an inflorescence, clustered together so as to appear as a single fruit' > la totalità; persone o animali o cose radunate insieme 'the whole; people, animals, or things gathered together'. This is due to the fact that the prominent semantic feature of this word is not the notion of fruit, but the fact that it refers to a group of entities, working as a collective noun. For this reason, the user does not find in our mapping ACINO and GRAPPOLO D'UVA under the notion of fruit, but in two different paths of the IWN network.

6. Conclusion

This paper presents a case study on mapping the ALT to IWN, focusing on the agricultural domain. We discussed the manual mapping process, our methodological choices and associated challenges. By releasing this dataset, we show how a highly structured and rich dialectological resource such as ALT can be enhanced and enriched by adhering to the Open Data principles and making it interoperable with other computational resources. Building on this work, we made some preliminary considerations on how the ALT-WordNet dataset can be exploited to closer inspect the concepts related to agriculture, and make a tentative conceptual grouping based on the IWN architecture. Future work will focus on exploiting the ILI (Interlingual Index) codes in IWN (Bartolini and Quochi, to appear) to project the ALT concepts onto the English WordNet, and allow for cross-linguistic studies.

7. Author contributions

PM manually mapped the agricultural domain to Italian WordNet and performed the evaluation and analysis of results. PM wrote sections 3, 4, 5, 6, and the Appendix, and contributed to section 1. SM provided overall supervision throughout the mapping process and manuscript preparation, including review and editing, and wrote sections 1 and 2.

8. Acknowledgements

This work was supported by the following projects: the NRRP Project PE 0000020 “Cultural Heritage Active Innovation for Sustainable Society” - CHANGES, CUP B53C22003890006, NRP Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU, within the activities of Spoke 3; the project “Una risorsa integrata del patrimonio linguistico-culturale toscano: l’Atlante Lessicale Toscano multimediale - multiALT” (’An Integrated Resource of Tuscan Linguistic and Cultural Heritage: the Multimedia Tuscan Lexical Atlas - multiALT’) funded by Regione Toscana (PROGETTI FSE+ 2021-2027) with the financial support of Accademia della Crusca.

9. Bibliographical References

Massimo Cerruti, Lorenzo Ferrarotti, Stefano Fiori, Matteo Rivoira, and Barbara Turchetta. 2025. Il progetto DigitALI: problemi e prospettive della digitalizzazione dell’Atlante Linguistico Italiano. *Linguistik online*, 137(5):37–67.

Beatrice Colcuc. 2020. La geolinguistica digitale e le sfide lessicografiche nell’era delle digital humanities: l’esempio di VerbaAlpina. In *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l’Informatica Umanistica*, pages 74–81, Milan, Italy. Associazione per l’Informatica Umanistica e la Cultura Digitale.

Nella Cucurullo, Simonetta Montemagni, Matilde Paoli, Eugenio Picchi, and Eva Sassolini. 2006. Dialectal resources on-line: the ALT-web experience. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).

Sebastiana Cucurullo, Simonetta Montemagni, Rubino Saccoccio, and Eva Sassolini. 2025. The Challenge of Obsolescence of Digital Archives in Cultural Heritage. A Case Study. *2025 IEEE 8th Congress on Information Science and Technology (CiSt)*, pages 506–512.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Hans Goebl. 2020. Presentazione delle due parti (ALD-I e ALD-II) dell’Atlante linguistico del ladino dolomitico e dei dialetti limitrofi. *Romance Philology*, 74:245–265.

Johann Mattis List Hans Geisler, Robert Forkel. 2021. A digital, retro-standardized edition of the tableaux phonétiques des patois suisses romands (tppsr). In André Thibaut, Matthieu Avanzi, Nicolas Lo Vecchio, and Alice Millour, editors, *Nouveaux regards sur la variation dialectale*, pages 13—-36. Editions de Linguistique et de Philologie, Strasbourg.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2393–2400, Portorož, Slovenia. European Language Resources Association (ELRA).

Eleonora Litta, Marco Carlo Passarotti, Valerio Basile, Cristina Bosco, Andrea Di Fabio, and Paolo Brasolin. 2025. Liita: a knowledge base of interoperable resources for italian. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 130–135.

Michele Loporcaro, Stephan Schmid, Chiara Zanini, Diego Pescarini, Giulia Donzelli, Stefano Negrinelli, and Graziano Tisato. 2021. Ais, reloaded: A digital dialect atlas of italy and southern switzerland. In André Thibaut, Matthieu Avanzi, Nicolas Lo Vecchio, and Alice Millour, editors, *Nouveaux regards sur la variation dialectale*, pages 111—-136. Editions de Linguistique et de Philologie, Strasbourg.

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking valency frames and wordnet synsets in the lila knowledge base of linguistic resources for latin. In *Further with Knowledge Graphs*, pages 16–28. IOS Press.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Eugenio Picchi, Simonetta Montemagni, and Lisa Biagini. 2001. DBT-ALT: A System for Storing and Querying the Data of the Atlante Lessicale

Toscano (ALT). *Dialectologia et Geolinguistica (DiG)*, 9:85–103.

Graziano Tisato. 2019. Acquisizione dell'intero AIS (Sprach- und Sachatlas Italiens und der Südschweiz). In Duccio Piccardi, Fabio Ardolino, and Silvia Calamai, editors, *Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale*, number 6 in Studi AISV, pages 131–153. Officinaventuno.

Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, and Barend Mons. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

10. Language Resource References

Roberto Bartolini and Valeria Quochi. to appear. ItalWordNet OMW. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli".

Giacomelli, Gabriella and Agostiniani, Luciano and Bellucci, Patrizia and Giannelli, Luciano and Montemagni, Simonetta and Nesi, Annalisa and Paoli, Matilde and Poggi Salani, Teresa. 2000. *DBT-ALT: Atlante Lessicale Toscano in CD-rom*. Lexis Progetti Editoriali.

Monica Monachini, Claudia Soria, and Antonio Toral. 2016. ItalWordNet kyoto. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli".

Montemagni, Simonetta and Picchi, Eugenio and Cucurullo, Sebastiana and Paoli, Matilde and Sassolini, Eva. 2006. *ALT-WEB: l'Atlante Lessicale Toscano in rete*. Istituto di Linguistica Computazionale "Antonio Zampolli", CNR.

Roventini, Adriana and Marinelli, Rita and Bertagna, Francesca. 2016. *ItalWordNet v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli".

Resources

We release the ALT-WordNet dataset and associated scripts in the following GitHub repository `https://github.com/paoma370/ALT-Web-to-WordNet.git`.

Appendix

Question Id	Question text (IT)	Question text (ENG)
87	Bosco di castagni	Chestnut grove
88	Bosco di castagni giovani	Young chestnut grove
89	Pollone di castagno	Chestnut shoot
90	Terriccio di castagno	Chestnut soil
91	Riccio della castagna	Chestnut burr
92	Buccia interna della castagna	Inner skin of the chestnut
93	Castagne vuote contenute nel riccio	Empty chestnuts in the burr
94	Seccatoio di castagne	Chestnut drying house / tool
115	Grappolo d'uva	Bunch of grapes
117	Parte del grappolo d'uva	Part of a bunch of grapes
118	Doppio grappolo	Double bunch (of grapes)
119	Raspo, grappolo senza acini	Grape stalk (bunch without grapes)
121	Viticcio	Tendril (of the vine)
122	Chicco d' uva	Grape
124	Sostegno morto della vite	Stake for the vine
125	Filare	Row (of vines)
126	Raccogliere le olive	To harvest olives
127	Frantoio	Olive oil mill / olive-press
128	Segale	Rye
129'	Mietere	To reap
129"	Mietitura	Reaping / harvest
130	Fascio di spighe legate sul campo	Sheaf (of ears of grain)
133	Palo del pagliaio	Haystack pole
134	Trebbiare	To thresh
135	Resti del grano nel campo	Remains of grain in the field (stubble)
135a	Resti del granturco nel campo	Remains of maize in the filed (corn stubble)
137	Pula	Chaff
139	Pagliuzza	Straw (small pieces)
140'	Vagliare il grano	To winnow grain
140"	Vaglio	Sieve
161	Vomere	Ploghshare
162	Bigoncia	Large wooden tub (for grapes/wine)
163	Grossa cesta rotonda per l'erba	Large round basket for grass
165c	Come si chiama il cestino per le olive?	What is the basket for olives called?
167	Orcio per l'olio	Oil jar
169	Cercine	Head pad (for carrying loads)
517	Scegliere	To select, to sort
519B	Raccogliere	To gather, to harvest

Table A1: Questions in the agriculture macro-class.

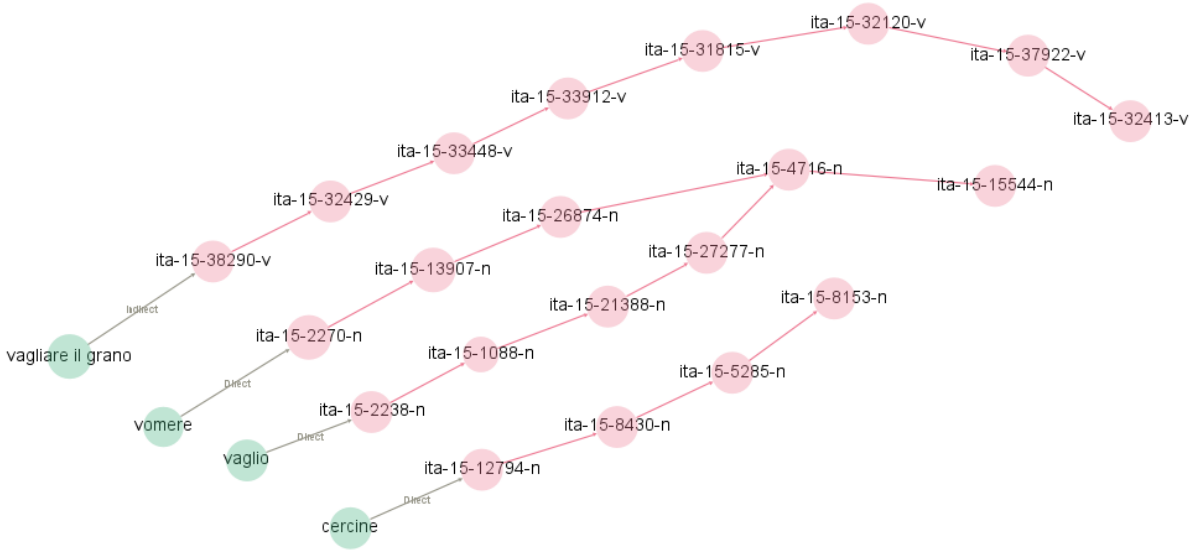


Figure A1: Combined WordNet subgraph for the concepts VAGLIARE IL GRANO, VOMERE, VAGLIO, and CERCINE, from the micro-class TOOL, for comparison with Figure 1. Green nodes represent the ALT concepts, while pink nodes represent IWN synsets. Coloured edges represent hierarchical relations between synsets (hypernymy). First- and second-level mapping are represented by grey edges, respectively indicated by the labels 'Direct' and 'Indirect'. As explained in section 5.2 in the main text, edges labeled as 'Indirect' are counted as one additional hierarchical step in the analysis. The concept described by the synsets in Figure is shown in Table A2.

synset_id	gloss (IT)	gloss (EN)
ita-15-1088-n	arnese atto all'esecuzione di determinate operazioni proprie di un'arte, di un mestiere o di una tecnica	tool used to perform specific operations typical of an art, craft, or technique
ita-15-12794-n	panno avvolto sulla testa per appoggiarvi brocche o ceste	cloth wrapped around the head to support jugs or baskets
ita-15-13907-n	piastra, lastra molto sottile	plate, very thin sheet
ita-15-15544-n	ciò che esiste	that which exists; entity
ita-15-21388-n	arnese o dispositivo atto a compiere determinate operazioni	tool or device designed to perform specific operations
ita-15-2238-n	attrezzo per setacciare	tool for sifting
ita-15-2270-n	lama dell'aratro che rovescia la zolla tagliata	plough blade that turns over the cut sod
ita-15-26874-n	pezzo di materiale rigido, di poco spessore e in genere rettangolare	piece of rigid material, thin and generally rectangular
ita-15-27277-n	tutto ciò che gli uomini producono	everything produced by human beings
ita-15-4716-n	oggetto materiale, concreto	material, concrete object
ita-15-5285-n	una parte di un tutto	a part of a whole
ita-15-8153-n	ciascuna delle porzioni o degli elementi in cui è diviso un tutto	each of the portions or elements into which a whole is divided
ita-15-8430-n	pezzo di tessuto	piece of fabric

Table A2: Synsets with Italian gloss and English translation from Figure A1.

A CLDF-Compliant Lexical Database for Modern Greek Dialects: Resource Design and Dialectometric Analysis

Stavros Bompolas[1,2], Natalia Chousou-Polydouri[2], Manuela Genitsaridi[2,3], Danae Karatzanou[2,3], Georgios Kostopoulos[2,3], Elena Anagnostopoulou[2,3], Dimitra Melissaropoulou[2,4]

[1]Archimedes/Athena RC, [2]Institute for Mediterranean Studies/Foundation for Research and Technology, [3]University of Crete, [4]Aristotle University of Thessaloniki

[1]Artemidos 1, 15125, Marousi, Greece, [2]Nikiforou Foka 130, 74131, Rethymno, Greece, [3]Department of Philology, 74100, Rethymno, Greece, [4]School of Italian Language and Literature, 54124, Thessaloniki, Greece

Correspondence: s.bompolas@athenarc.gr

Abstract

This paper presents the first CLDF database that systematically documents lexical variation across 36 Modern Greek varieties (including Standard Modern Greek). The dataset aligns 14,378 lexical items over 345 concepts, links varieties to stable identifiers (Glottocodes), introduces a Greek-specific concept list, and maps meanings to standardized Concepticon concept sets, enabling interoperability and reproducible workflows. To assess whether the database preserves a meaningful dialectological signal, we conduct a dialectometric analysis by computing feature-sensitive string distances over IPA transcriptions and applying hierarchical clustering. The resulting similarity structure recovers major macro-divisions—most notably a broad Northern vs. Southern partition among Koine-descended mainland varieties—and isolates peripheral groups with distinct historical trajectories (e.g., Asia Minor, Italiot, Tsakonian). The database provides scalable infrastructure for quantitative dialectology, comparative Greek linguistics, and dialect-aware language technology.

Keywords: Modern Greek dialects, lexical variation, lexical database, CLDF, dialectometry

1. Introduction

Lexical databases constitute core infrastructure for research in linguistic typology, historical linguistics, and Natural Language Processing (NLP). The adoption of structured and interoperable standards—most prominently Cross-Linguistic Data Formats (CLDF)—has significantly improved the transparency, comparability, and long-term reusability of lexical data (Forkel et al., 2018b). These advances have enabled large-scale cross-linguistic comparison and computational analysis at an unprecedented scale. Yet this progress has been uneven: while variation *across* languages is increasingly well documented, systematic dialectal variation *within* languages remains substantially underrepresented. This imbalance reflects a persistent standard-centric bias in resource development, whereby non-standard and regional varieties are marginalized, normalized, or reduced to peripheral deviations from an assumed "main" language.

Modern Greek (MG) exemplifies this asymmetry. Although Standard Modern Greek (SMG) is included in several major lexical and typological databases, dialectal varieties receive only limited coverage. Large cross-linguistic lexical databases include at most a small number of Greek varieties (e.g., IE-Cor includes four MG varieties, Heggarty et al., 2023), while typological resources typically encode a single standard-based profile (e.g., Grambank includes only SMG, Skirgård et al., 2023; WALS includes SMG and Cypriot Greek, Dryer and Haspelmath, 2013). At the same time, Greek dialectology has produced extensive descriptive scholarship, including dialect dictionaries, grammars, and regional atlases. Yet digital infrastructures that systematize and integrate this material remain scarce. As a result, despite a rich dialectological tradition, there is currently no large-scale, concept-aligned, computationally reusable database that documents lexical variation across the breadth of the MG dialect continuum in a format compatible with contemporary lexical infrastructures and NLP workflows.

In this paper, we address this infrastructural gap by presenting a CLDF-compliant lexical database documenting dialectal variation across 36 MG varieties (including SMG), aligned over 345 concepts and comprising 14,378 lexical items. The dataset transforms dispersed dialect material into a structured, interoperable resource that renders intra-language lexical variation computationally tractable, while providing a foundation for future multi-level extensions (phonology, morphology, and syntax). Beyond introducing the resource, we evaluate whether it encodes a meaningful dialectological signal through dialectometry, applying feature-sensitive distance measures and hierarchical clus-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 287–299
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tering to test whether aggregate lexical–phonetic similarity recovers groupings anticipated in the dialectological literature. Concretely, the paper:

1. introduces the first CLDF-compliant lexical database covering a broad sample of MG dialects;
2. documents the methodological and modeling decisions required to represent a dialect continuum within an interoperable framework; and
3. offers a dialectometric evaluation that both validates the internal coherence of the dataset and provides a data-driven reassessment of traditional dialect classifications.

2. Related Work

Our work sits at the intersection of (a) cross-linguistic lexical infrastructures grounded in concept-list methodology, (b) emerging efforts to develop computationally reusable resources for dialect continua and closely related varieties, and (c) MG dialect lexicography, where lexical documentation remains fragmented and infrastructurally underdeveloped. We therefore review prior work along these three dimensions.

2.1. Cross-Linguistic Lexical Infrastructures

Linguists have long relied on standardized concept lists to compare the lexicons of different languages (e.g., Swadesh lists, Swadesh, 1955; the Leipzig–Jakarta list, Haspelmath and Tadmor, 2009a; among others). Building on this tradition, large-scale databases now provide comparative wordlists across hundreds or thousands of languages. For example, the Automated Similarity Judgment Program (ASJP) maintains 40-item wordlists in a unified phonetic encoding for most of the world's languages (Wichmann et al., 2025). The World Loanword Database (WOLD) provides "mini-dictionaries" of approximately 1,000–2,000 entries for 41 diverse languages, with each entry annotated for loanword versus inherited status (Haspelmath and Tadmor, 2009b). Like ASJP, WOLD relies on a fixed meaning list (1,460 concepts derived from the Intercontinental Dictionary Series; Key and Comrie, 2023), enabling systematic cross-language comparison of lexical structure and borrowing patterns. In addition to global comparative resources, there are family-specific lexical databases designed for particular lineages, such as IE-CoR for the Indo-European family (Heggarty et al., 2023) and LA80 for a set of North-West Bantu A80 languages (Vermeir et al., 2024).

More recently, efforts have focused on improving interoperability and standardization. The Cross-Linguistic Data Formats (CLDF) initiative proposes simple, tabular standards for lexical and structural datasets (Forkel et al., 2018b), along with associated tools for validation, processing, and reproducible dataset curation (Forkel et al., 2018a; Forkel and List, 2020; Kaiping et al., 2022). Built on CLDF, Lexibank aggregates dozens of published wordlist datasets into a unified, FAIR-compliant repository (List et al., 2022). In its current form, Lexibank covers over 2,000 language varieties drawn from roughly 100 source datasets. Crucially, it harmonizes transcription practices and links entries to standardized identifiers (e.g., Glottocodes for languages and Concepticon IDs for meanings), thereby ensuring interoperability and reproducibility. Complementing these efforts, Concepticon maps tens of thousands of concept labels from diverse published lists onto a few thousand standardized concept sets (List et al., 2025), facilitating cross-dataset comparison even when original glossing conventions differ.

Taken together, these infrastructures have transformed lexical typology and historical comparison into transparent and computationally tractable enterprises. However, their primary unit of organization remains the language. Typically, one representative lexical item per concept is recorded for each variety. As a result, while these resources excel at cross-language comparison, they are not designed to capture systematic microvariation within a single language, where multiple competing realizations per concept and fine-grained sociolinguistic metadata may be central.

2.2. Lexical Resources for Dialects

Resources specifically designed for closely related varieties (dialects, regional standards, or "similar languages") are comparatively rare. Traditional dialectology has produced extensive linguistic atlases (see Lameli, 2009; Kretzschmar, 2017) and regional glossaries or dialect dictionaries (see Moulin, 2009; Van Keymeulen, 2017), which document lexical and structural variation at fine geographical resolution. However, these works are typically published as printed maps, volumes, or static digital repositories. Even when digitized, data are often distributed across heterogeneous formats and lack standardized encoding. Coverage may be uneven, transcription practices inconsistent, and metadata incomplete. Consequently, although dialect atlases and glossaries are rich in empirical detail, they are rarely structured in a way that supports systematic, item-by-item comparison across many varieties or straightforward computational reuse.

This underrepresentation persists even within contemporary interoperable infrastructures. For example, in Lexibank—currently comprising approximately 75 curated lexical datasets—only 14 ex-

plicitly target dialectal or closely related varieties, covering just 87 varieties out of more than 2,029 included overall. Moreover, these datasets typically focus on relatively small and geographically restricted areas. This reflects a broader pattern: although current infrastructures technically support dialect-level encoding, dialect continua remain substantially underrepresented in practice.

The relative scarcity of such resources underscores the methodological and infrastructural gap that persists for dialect continua—including MG.

2.3. Lexical Resources for Modern Greek Dialects

MG dialectology has a long and productive tradition, including detailed monographs, dialectal dictionaries, and grammatical descriptions of dialects (Tzitzilis, 2000; Katsouda, 2024). Nevertheless, dialect documentation has typically proceeded on a per-variety basis, resulting in isolated lexica or grammatical studies rather than integrated, cross-dialect resources (for a relevant discussion, see Trudgill, 2003).

More recent digital initiatives focus on specific varieties or limited geographical areas—for example, work on Cappadocian Greek (Melissaropoulou et al., 2022; ILIK, 2024). Other efforts remain under development and are not yet released as openly-licensed, interoperable datasets suitable for systematic comparison and computational reuse (ILNE, 2026). As a result, despite the richness of dialectological scholarship, there is currently no large-scale, computationally reusable database that systematically documents variation across the breadth of MG dialects in a format suitable for structured comparison and NLP-oriented reuse.

Cross-linguistic lexical infrastructures occasionally include MG data, but typically restrict coverage to the standard variety. A partial exception is IE-CoR (Heggarty et al., 2023), which incorporates several Greek varieties, including Pontic, Cappadocian, Cypriot, Italiot, Peloponnese Tsakonian, and Propontis Tsakonian. Similarly, the Global Lexicostatistical Database (Starostin, 2011) documents SMG alongside Peloponnese Tsakonian, Pharasa Greek, and Cappadocian Greek (Aravan). In both cases, coverage remains limited and does not aim at systematic intra-language dialect comparison; moreover, some entries rely on abstracted or supra-dialectal forms—for instance, Cappadocian data in IE-CoR are derived from Dawkins' glossary (Dawkins, 1916), which presents normalized lemmas rather than systematically differentiated subvarieties.

In practice, MG dialectal data remain dispersed across independent projects and publications, without unified alignment, consistent metadata standards, or interoperable encoding. This gap constrains both dialectological research and computational applications.

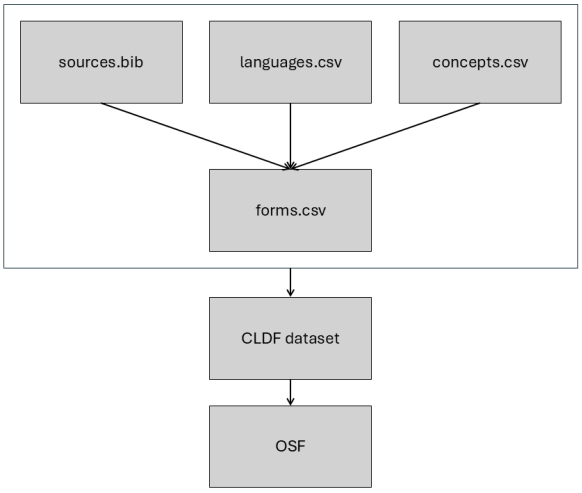


Figure 1: Structure of the CLDF dataset.

Our work addresses this gap by compiling a multi-dialectal MG lexical database in CLDF format, explicitly aligned by concept and designed for computational reuse. In doing so, we integrate Greek dialect data into contemporary lexical infrastructures and create a foundation for dialectometric analysis, NLP diagnostics, and future expansion toward multi-level linguistic documentation.

3. Resource Design

Our database conforms to the Cross-Linguistic Data Formats (CLDF) standard (Forkel et al., 2018b). The current release is structured around the standard CLDF Wordlist module and consists of four core files:

- concepts.csv
- forms.csv
- languages.csv
- metadata.json

In addition, a `sources.bib` file contains the bibliographic references associated with each entry.

The `metadata.json` file defines the dataset structure and explicitly specifies how the tables are interrelated (e.g., linking forms to concepts and language varieties). CLDF is modular and extensible: additional columns or new tables can be added without compromising compliance with the standard. This flexibility is essential for future extensions of the database beyond lexical data. All data are publicly available via OSF.

The core CSV files are described below, highlighting their principal columns and coding decisions. The overall CLDF architecture is presented in Figure 1.

3.1. Dialect Coverage and Selection Criteria (languages.csv)

The `languages.csv` table contains one row per variety (doculect), each identified by a unique ID. The table includes structured metadata such as:

- Glottocode (serves as a unique ID)
- Name
- Longitude
- Latitude
- Locality
- ISO code (if available)
- Phonetic conversion notes
- Comment (free-text)

The current release covers 36 MG varieties (including SMG) spoken in and outside Greece in the 20th century (Figure 2).

A central design decision concerned the selection of varieties to be included in the database. Our sampling strategy aimed to balance three criteria:

1. **Geographical representativity**: coverage of the major dialect areas across the MG continuum, including mainland, insular, and Asia Minor varieties.
2. **Dialectological representativity**: inclusion of members from all traditionally recognized macro- and micro-groupings in Greek dialectology, ensuring comprehensive coverage of the established classificatory landscape.
3. **Documentary adequacy**: availability of sufficient reference materials to support reliable data extraction and future extensibility (including to other levels of linguistic analysis, thereby ensuring cross-level comparability).

Given these criteria, we prioritized varieties for which both lexical and grammatical documentation are available. Because the long-term goal is to expand the database to additional linguistic levels (e.g., phonology and morphosyntax), the inclusion of both types of documentation ensures structural comparability and facilitates future development. As a result, some historically important varieties that are lexically documented but lack sufficient grammatical description (e.g., the Greek of Constantinople) were not included in the present release.

Additionally, to account for intra-dialectal diversity, multiple varieties were included for certain groups (e.g., Cappadocian, Pontic, Italiot Greek). This serves two purposes: (a) to capture significant internal variation often resulting from contact-induced change; and (b) to avoid reliance on supra-dialectal or artificially normalized lemmas that obscure local variation.

Glottocodes. ISO 639-3 codes are limited in dialect coverage and often represent abstract language groupings. As part of this project, we systematically mapped, updated, or created Glottocodes for the varieties in our sample (Hammarström and Forkel, 2022). Each doculect is identified by a Glottocode and Glottolog name.

Many MG dialects are inconsistently represented—or entirely absent—from widely used language catalogues. This complicates dataset linking, reproducibility, and interoperability. Establishing stable identifiers provides:

- Unambiguous reference to dialect entities
- Improved dataset merging and comparison
- Durable cross-resource integration
- Support for cumulative work (e.g., adding grammatical features or linking corpora)

We mapped existing Glottocodes where possible, refined imprecise legacy entries, and introduced new identifiers for previously uncoded varieties.[1]

3.2. Definition of the Concept List (concepts.csv)

The database is organized around a concept list of 345 concepts covering core vocabulary across parts of speech (nouns, verbs, adjectives, and functionally salient items).

The `concepts.csv` file contains one concept per row with associated metadata:

- ID
- Concept
- Comment (precisions on the intended meaning and contexts)
- Concepticon ID
- Concepticon Name
- Concepticon Comment (precisions regarding the Concepticon mapping)

The concept list was constructed by combining and adapting established concept inventories widely used in historical linguistics, especially in research on the Indo-European family (e.g., Dyen et al., 1992; Ringe et al., 2002; Scarborough, 2019). Since there is, to our knowledge, no dedicated concept list specifically designed for the Greek branch or for systematic comparison across MG dialects, the present study also represents an initial step in that direction. In the absence of a Greek-specific benchmark, we used these earlier concept lists as a starting point and adapted them to the lexical and dialectological particularities of Greek as follows:

- merging overlapping concepts,
- resolving differences in semantic granularity,
- removing redundant, sparsely attested, or diagnostically uninformative items, and
- adding or refining concepts that are diagnostically relevant for Greek dialect differentiation.

[1] New and updated Glottocodes will be included in the upcoming Glottolog release (v5.3), but are already available in the Glottolog GitHub repository: https://github.com/glottolog/glottolog.

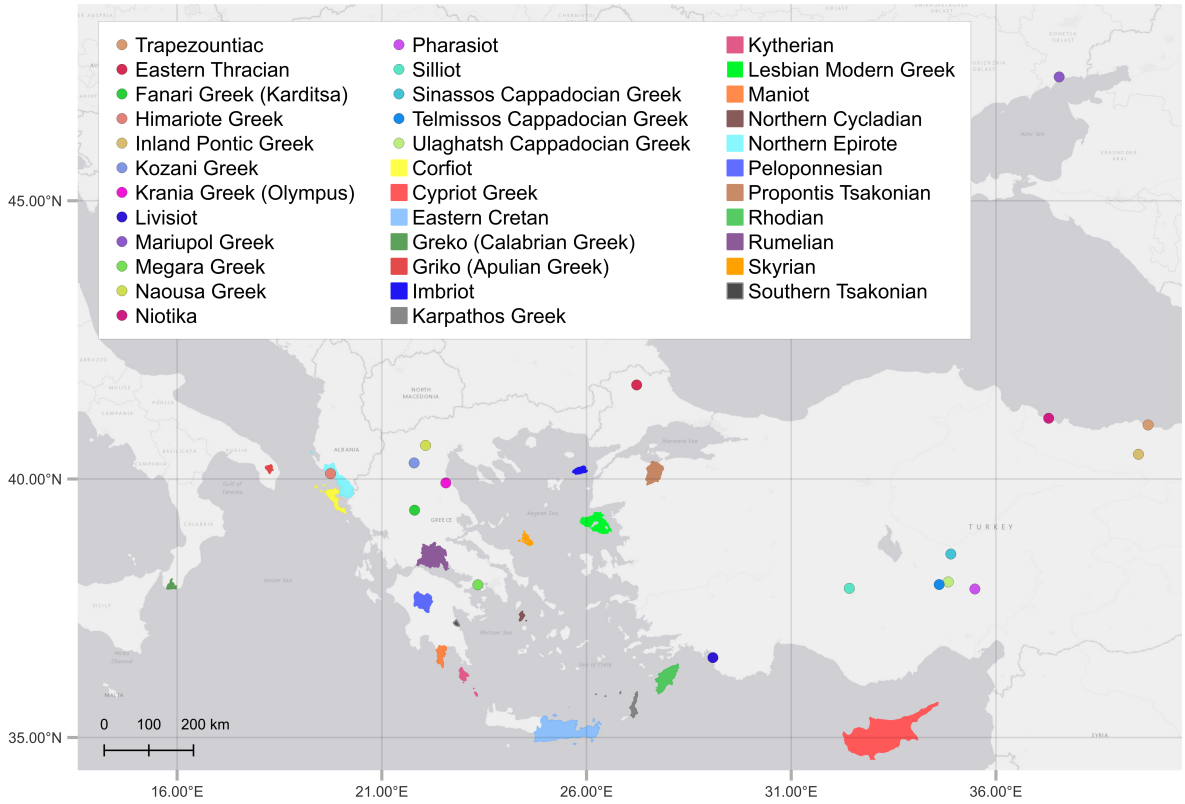


Figure 2: Geographic distribution of the sampled MG varieties represented in the database (excluding SMG).

Items were excluded when they were poorly documented, overly culture-specific, or unlikely to contribute to meaningful cross-dialect contrast. For example, collective items such as GRAIN and CATTLE were removed due to inconsistent documentation across dialect sources. We also eliminated pairs that are etymologically related within the Greek branch and therefore unlikely to yield independent contrastive evidence (e.g., SISTER VS. BROTHER, HE VS. THEY, HE VS. THIS).

Conversely, we split concept pairs that are not etymologically related in Greek in order to avoid masking dialectal differentiation. Nominal and verbal pairs such as SLEEP VS. TO SLEEP and VOMIT VS. TO VOMIT were treated as distinct concepts. We further refined semantically broad items where dialects exhibit systematic lexical differentiation, including FIGHT (WITH SOMEONE) VS. FIGHT (IN WAR), male vs. female goat terminology, and SCRATCH (OF CAT) VS. SCRATCH (AN ITCH). In addition, OLD MAN and OLD WOMAN were introduced as distinct concepts, separate from OLD as an adjectival property applicable to objects.

This approach preserves comparability with established historical-linguistic practice while making it possible to capture distinctions that are especially relevant for dialect-level comparison within Greek. It prioritizes broad attestation across dialect sources, semantic stability, diagnostic value for dialect differentiation, and compatibility with cross-linguistic practice.

Crucially, the list is aligned with Concepticon, linking each entry to standardized concept sets. This alignment ensures cross-dataset comparability, reproducibility, and interoperability with other CLDF-based resources and typological datasets.

3.3. Lexical Entries (forms.csv)

Lexical entries were compiled from dialect dictionaries, glossaries, grammars, published texts, and verified digital databases (including IE-Cor). Where discrepancies were identified, forms were systematically checked against primary sources or dialect speakers when available.

The `forms.csv` table contains one row per dialectal form. Each entry links a specific dialect form to a concept and a language variety. Core fields include:

- ID: Unique identifier
- Language_ID: External reference to the *ID* column of the `languages.csv` table

- Concept_ID: External reference to the *ID* column of the `concepts.csv` table
- Orthographic_form: Attested spelling in the source
- Phonetic_form: IPA transcription
- Phonemic_form: Phonemic representation, where available
- Segments: Segmented IPA representation
- Gender: Grammatical gender (for nouns)
- Translation: Disambiguating gloss and additional meanings, as they appear in the source
- Variations: Attested intra-dialectal variants
- Source: Bibliographic reference, including page numbers
- Source_comment: Notes from the original source (e.g., etymology, usage restrictions, loan status)
- Curator_comment: Editorial comments added by the dataset curator

Orthographic transcriptions follow strictly the spelling conventions of the original sources, with no normalization applied, so that forms can be located directly using the cited references. The only systematic adaptation concerns the removal of polytonic diacritics, which do not encode dialect-specific phonetic information and mainly reflect older printing conventions rather than pronunciation.

The IPA layer required substantial editorial work. Greek dialects lack a unified orthographic standard, and no comprehensive orthography–IPA mapping exists for all varieties. Manolessou et al. (2012) provide the most systematic attempt toward such standardization and serve as a key reference point, particularly for older descriptions, although their coverage remains far from exhaustive. Moreover, some dialect sources deliberately avoid representing phonetic particularities and instead employ SMG orthography to emphasize lexical rather than phonetic differentiation. Therefore, we adopted a multi-step approach: (a) for varieties covered in Manolessou et al. (2012), we followed their transcription guidelines; (b) where IPA or IPA-like transcriptions were provided in the sources, these served as the primary basis; (c) in other cases, IPA forms were established on the basis of grammatical descriptions and general dialectological studies. In such cases, greater divergence between orthographic and phonetic representations may reflect the original editorial choices of the sources. Where necessary, transcription conventions specific to individual varieties are documented in `languages.csv`.

The `Segments` field is derived directly from the `Phonetic_form` field. In the current release, segmentation follows the IPA string and treats as single segments only those consonant clusters that are explicitly marked as unitary by means of a tie bar. At present, this applies only to the dialectal affricates [t͡ʃ] and [d͡ʒ], which in MG dialects typically derive from earlier /k/ and /g/, respectively, before /i, e/. By contrast, clusters such as [ts] and [dz] are currently segmented into two symbols. This choice is deliberate, since their status as single phonemes in MG remains disputed and available research on their phonological behavior is not conclusive enough to justify treating them uniformly as single segments in the dataset (for an overview, see Arvaniti, 2007, 114–117). One limitation of this approach is that dialectal sequences such as [ts] and [dz] may still be segmented into two symbols even in cases where they originate from a single phoneme. This issue will be addressed more systematically in future versions of the dataset.

3.4. Data Format and Interoperability

The database is released in CLDF, ensuring:

- Interoperability with existing lexical comparison and dialectometric tools
- Transparent separation of forms, concepts, and metadata
- Integration with infrastructures such as Concepticon and Lexibank
- Long-term extensibility as new linguistic layers are added

CLDF enables the resource to function not only as a standalone dataset but as an extensible component in a broader ecosystem of dialect-aware NLP and comparative linguistic workflows.

3.5. Automated Validation and CLDF Generation

Throughout the coding process, `pycldf` (Forkel et al., 2018a) and `lexedata` (Kaiping et al., 2022) were used to:

- Transform working spreadsheets into CLDF format
- Validate IPA transcriptions
- Check references consistency
- Detect structural errors and warnings
- Split the IPA representations into segments

After validation, the dataset was exported into its final CLDF form. All files are publicly available via OSF.

3.6. Dataset Coverage and Descriptive Statistics

The dataset's completeness varies by variety, reflecting differences in documentary depth. Across the 36 doculects, concept coverage ranges from 53/345 (15.4%) at the low end to 344/345 (99.7%) at the high end, with a median coverage of 85.4% (294.5 concepts). The overall distribution of concept coverage across varieties is shown in Figure 3. In addition to coverage, the database captures substantial intra-dialectal lexical variation: the mean

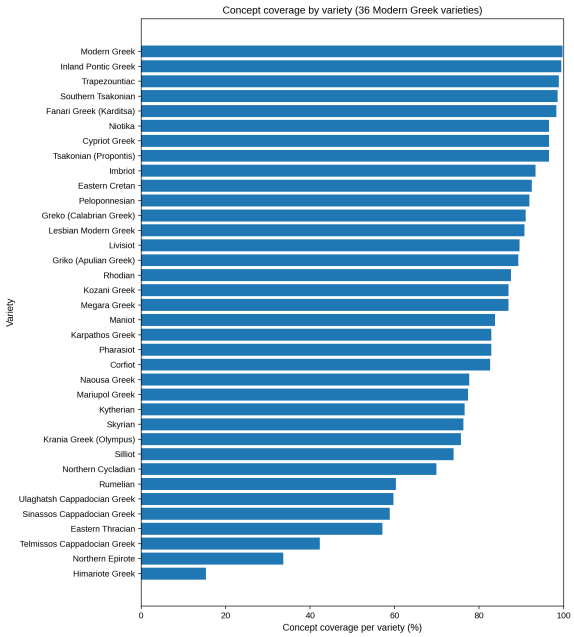


Figure 3: Concept coverage by variety.

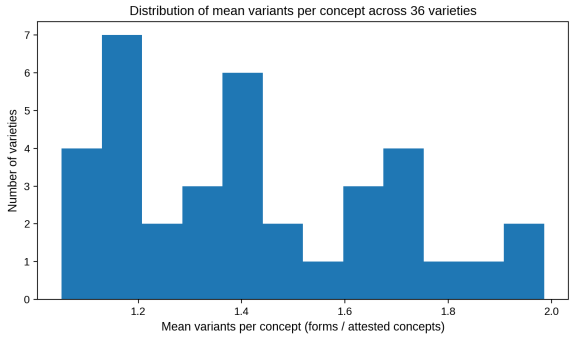


Figure 4: Mean lexical items per concept.

number of lexical items per attested concept ranges from 1.05 to 1.99 across varieties (median 1.37), indicating that many doculects record multiple competing realizations for the same concept. The distribution of this "variation density" measure is shown in Figure 4. Together, these statistics quantify both (a) the dataset's broad comparability across the dialect continuum and (b) its ability to represent microvariation within individual varieties.

4. Dialectometric Analysis

We use the lexical database introduced above to (a) evaluate whether it encodes a dialectologically plausible signal and (b) explore aggregate, data-driven (Nerbonne, 2009) groupings of MG varieties that can be compared to established dialectological classifications.

While dialectometry has substantially advanced modern dialectology (cf. Wieling and Nerbonne, 2015; Goebl, 2017; Heeringa and Prokić, 2017; Nerbonne and Wieling, 2017; Nerbonne et al., 2021; for an overview of computational dialectology with Greek-oriented examples, see Bompolas, 2023, Ch. 2), MG dialectology has only partially benefited from these developments. Existing computational studies have typically focused on restricted geographic areas (e.g., for inner Asia Minor Greek, see Bompolas, 2023; for the Pontic–Azov area, see Kisilier and Solliec, in press) or include only a small number of varieties (cf. Chatzikyriakidis et al., 2026), largely due to infrastructural constraints: the absence of a comprehensive dialect atlas covering the full dialect continuum and the limited availability of large, digitized dialectal corpora or machine-readable dictionaries. Notably, many of the challenges identified by Trudgill (2003, 46–48) for the classification of MG dialects remain relevant, as resource fragmentation and digitization gaps continue to limit quantitative work across broad samples.

Against this background, we provide a dialectometric evaluation grounded in our concept-aligned lexical database spanning a wide range of MG varieties. To our knowledge, no previous dialectometric study has combined this breadth of MG varieties with a concept-aligned, computationally-reusable lexical dataset. We next outline our workflow, which comprises (a) data extraction and parameterization from the CLDF release, (b) computation of lexical–phonetic distances, and (c) exploratory statistical analysis via clustering.

4.1. Data Preparation

For dialectometric analysis, we use the LED-A web application (Heeringa et al., 2023). The `forms.csv` table of the CLDF database was converted into the tabular format required by LED-A. Concretely, we constructed a data matrix in which (a) columns correspond to the 345 concepts, (b) rows correspond to the 36 varieties, and (c) cells contain IPA transcriptions. To maximize comparability, we rely exclusively on the IPA layer of the database, which represents attested forms in a consistent manner and thus provides a reliable basis for measuring dialect distances. When more than one lexical item (either variant or synonym) is attested for a concept in a given variety, all items are encoded within the same cell and separated by a slash. Missing data are left blank.

Since concept coverage varies across varieties (cf. Figure 3), distance computations are necessarily based on the subset of concepts attested for each pair of varieties. Consequently, comparisons involving sparsely documented varieties rely on fewer shared data points and may therefore yield less stable results than comparisons involving better documented varieties.

4.2. Lexical–phonetic Distance Measures

Dialectometric analysis transforms qualitative observations (lexical items, in our case) into a symmetric distance matrix in which each cell expresses the degree of linguistic difference between a pair of varieties. LED-A implements several string-based distance measures. We selected the distance measure that performed best under *local incoherence*, a diagnostic introduced by Nerbonne and Kleiweg (2007): geographically proximate varieties are generally expected to be linguistically more similar, and lower local incoherence indicates a distance measure that better fits this dialectological expectation (see also Heeringa et al., 2006).[2]

Among the available LED-A options, the `A&B sub. ≤ 1, indel ≤ 0.5` setting yielded the lowest local incoherence for our dataset. This corresponds to a feature-sensitive Levenshtein-style distance in which substitution costs depend on phonetic feature similarity (cf. Heeringa and Braun, 2003; Heeringa, 2004). Insertions and deletions are penalized with lower maximum cost than substitutions.[3] The output is a symmetric site × site distance matrix (36×36), which serves as input for clustering analysis.

4.3. Hierarchical Clustering

To explore structure in the distance matrix, we apply hierarchical clustering and visualize the results as a dendrogram (for its usage in dialectometry, see Prokić and Nerbonne, 2008; see also Nerbonne and Wieling, 2017, 401–403). We tested the linkage methods implemented in LED-A and selected the one that accounted for the most variance in our data. In our setting, UPGMA achieved the highest explained variance (71.1%).

4.4. Results and Comparison to Dialectological Classifications

Figure 5 presents the UPGMA dendrogram derived from feature-sensitive string distances. For interpretability, we highlight eight major clusters (color-coded in the figure).

At the highest level, the clustering recovers a broad *Northern* vs. *Southern* partition within the mainland Koine-descended varieties, consistent with accounts that emphasize the phonological

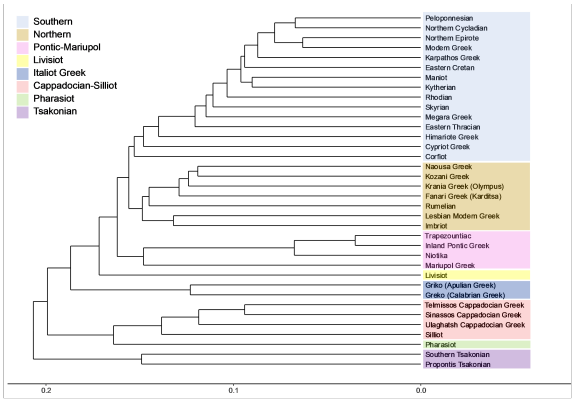


Figure 5: UPGMA dendrogram showing dialect similarity across 36 MG varieties; eight clusters are highlighted.

criterion of Northern vocalism (i.e., raising of unstressed /e, o/ and deletion of unstressed /i, u/) as a major diagnostic, originally proposed by Hadzidakis (1892) (see also Trudgill, 2003, 49–50 and references therein).

Several deviations from a purely geographic North–South split are also informative. In our sample, the geographically northern varieties Skyros and the Eastern Thracian variety (Saranta Ekklisies) pattern closer to the Southern cluster. Both have been described as *semi-northern* or as lacking the full set of Northern vocalism diagnostics (e.g., Psaltis, 1905, 12; Papadopoulos, 1926, 12,14; Katsouda, 2021, 53). Under a distance measure that captures cumulative lexical–phonetic similarity, their aggregate profiles align more closely with the Southern than with the Northern cluster; this is compatible with independent observations that they also share properties with southern varieties (e.g., Katsouda, 2021, 299–300 for Skyrian). Another exception to this geographically aligned partition concerns the varieties of southern Albania (Northern Epirote and Himariote), which do not exhibit Northern vocalism (Kyriazis and Spyrou, 2011, 179). For the remaining Northern varieties, the clustering broadly aligns with traditional expectations, and geographically proximate varieties tend to cluster together.

Within the Southern group, finer-grained subgroupings proposed in the literature are not consistently recovered (see Trudgill, 2003, 49–50 and references therein). This may reflect (a) the fact that the present analysis is based on aggregate lexical–phonetic distances rather than a small set of diagnostic features (e.g. Trudgill's classification relies on six phonological features), and (b) differences in how features are represented in the sources used here. For instance, some subgroup descriptions may be too coarse: Trudgill (2003)

[2] A limitation is that SMG is represented by the coordinates of Athens (as a proxy for its early-20th-century sociolinguistic center), which may slightly distort local incoherence for this single data point. Since this affects only one variety, we consider the diagnostic still informative for model selection.

[3] See the LED-A documentation for the implementation of feature string comparison: `https://www.led-a.org/docs/A&B.pdf`.

notes that Peloponnesian does not exhibit any of the six features used in his classification (pp. 58–59) and that Northern Cycladian is characterized only by tsitakism (p. 60). In our sources, however, both varieties also show velar palatalization, which plausibly contributes to their proximity in our aggregate distance-based clustering. In addition, lexical borrowing can influence similarity structure: Corfiot, for example, appears relatively peripheral within the Southern cluster, plausibly due to extensive Italo-Venetian lexical influence in the data.

Modern Greek (i.e., SMG) likewise patterns clearly with the Southern cluster. This placement is historically plausible, given the widely acknowledged contribution of the Ionian Islands and the Peloponnese to the formation of the Standard (following Horrocks, 2010; see also Trudgill, 2003, 58–59), though the Peloponnesian input has been debated (Pantelidis, 2001). In the dendrogram, SMG clusters closely with Peloponnesian but does not form an immediate sister grouping, reflecting phonetic differences in the dataset.

Beyond the mainland core, several peripheral clusters emerge in line with expectations about contact, relative isolation, and distinct historical trajectories. Within the Pontic—Mariupol Greek cluster, the dendrogram yields a coherent internal subdivision that accords with traditional groupings (e.g. the West vs. East distinction and further Inland/Trapezountiac differentiation; Triandaphyllidis, 1938, 288), while Mariupol Greek clusters closely with Pontic, in line with prior quantitative work suggesting their affinity (Kisilier and Solliec, in press). A particularly instructive case is Livisiot, which emerges as a peripheral branch rather than clustering tightly with either southern varieties or the inner Asia Minor group, contrary to some earlier proposals (see Andriotis, 1961, 11–15 for an overview). Italiot Greek forms a compact pair, reflecting shared contact history and long-term divergence from mainland varieties (cf. Horrocks, 2010, Ch. 14.2.3). Within inner Asia Minor, Cappadocian/Silliot and Pharasiot show a subdivision compatible with qualitative accounts (Manolessou, 2019, 30 and references therein) and earlier computational results (Bompolas and Melissaropoulou, 2023) based on shared phonological traits. Finally, Tsakonian (Propontis and Southern) forms the most divergent cluster, as expected given its distinct genealogical status within Greek and its well-known structural distance from Koine-descended varieties (Horrocks, 2010, Ch. 4.4.3).

Overall, the dialectometric evaluation supports two conclusions. First, the database encodes a clear dialectological signal: major macro-divisions and established subgroupings emerge from aggregate lexical–phonetic distances without being hard-coded into the dataset. Second, mismatches with traditional classifications are interpretable rather than arbitrary, highlighting cases where (a) dialect labels obscure internal heterogeneity, (b) a variety occupies an intermediate position with respect to canonical diagnostics, or (c) borrowing and contact influence similarity structure. These results demonstrate the value of concept-aligned lexical resources for computational dialectology and provide a baseline for future analyses incorporating additional linguistic levels as the resource expands.

5. Conclusion and Future Work

This paper introduced the first systematically aligned, CLDF-compliant lexical database designed to capture lexical variation across the MG dialectal continuum. The current release covers 36 MG varieties (including SMG), aligned over 345 concepts, and includes 14,378 lexical items, encoded with stable identifiers, explicit metadata, and a harmonized IPA layer to support interoperability and reproducible reuse. Alongside the resource description and the modeling decisions required to represent a dialect continuum in an interoperable framework, we evaluated whether the dataset preserves a meaningful dialectological signal. A dialectometric analysis based on feature-sensitive string distances and hierarchical clustering recovers major macro-divisions—most notably the broad Northern vs. Southern partition among Koine-descended mainland varieties—and isolates historically divergent peripheral groups.

Overall, the database provides infrastructure for quantitative dialectology, comparative Greek linguistics, and dialect-aware language technology, establishing a baseline for research on Greek microvariation. It can also support downstream computational tasks such as automatic cognate (List, 2012) and loan (List and Forkel, 2022) detection, as well as dialect-aware NLP (e.g., multi-dialect training and transfer learning, Lin et al., 2019; Faisal and Anastasopoulos, 2022).

We foresee four main directions for further development. First, we will extend the database beyond the lexical layer by adding grammatical features, moving toward a multi-level resource that captures phonological, morphological, and syntactic variation across MG varieties. Second, we plan to incorporate earlier stages of Greek to enable diachronic analyses linking historical change to present-day dialect differentiation. Third, we will enhance accessibility through a web-based interface (e.g., CLLD) alongside the archival CLDF release. Fourth, we will broaden the analytical framework beyond exploratory clustering to include additional dialectometric and phylogeny-oriented methods, systematically comparing quantitative results with established classifications.

6. Data Availability

The CLDF dataset and the corresponding matrix used for the dialectometric analysis are available on OSF: https://osf.io/t5avq/overview.

7. Acknowledgements

This work has been funded by the European Union (ERC, PhylProGramm, 101096554). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

We also express our sincere gratitude to Harald Hammarström for his invaluable assistance with the systematic assignment and revision of Glottocodes and for his support in ensuring their integration into Glottolog.

8. Bibliographical References

Nikolaos P. Andriotis. 1961. *The Idiom of Livisi in Lycia [in Greek]*. Centre for Asia Minor Studies, Athens.

Amalia Arvaniti. 2007. Greek Phonetics. The State of the Art. *Journal of Greek Linguistics*, 8(1):97–208.

Stavros Bompolas. 2023. *Computational dialectology in the linguistic varieties of Cappadocian, Pharasiot, and Silliot*. Ph.D. thesis, University of Patras, School of Humanities and Social Sciences, Department of Philology, Linguistics Section.

Stavros Bompolas and Dimitra Melissaropoulou. 2023. A dialectometric approach to inner Asia Minor Greek: Comparisons and associations between linguistic levels. *Digital Scholarship in the Humanities*, 38(4):1389–1403.

Stergios Chatzikyriakidis, Erofili Psaltaki, Dimitrios Papadakis, Erik Henriksson, and Veronika Laippala. 2026. Perplexity as a Metric for Dialectal Distance: A Computational Study of Greek Varieties. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 101–112, Rabat, Morocco. Association for Computational Linguistics.

Richard MacGillivray Dawkins. 1916. *Modern Greek in Asia Minor: a study of the dialects of Sílli, Cappadocia and Phárasa with grammar, texts, translations and glossary*. Cambridge University Press, Cambridge.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo. Type: Data set.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society*, 82(5):iii–132.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-Inspired Adaptation of Multilingual Models to New Languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Robert Forkel, Simon J. Greenhill, and Christoph Rzymski. 2018a. cldf/pycldf: pycldf. Zenodo.

Robert Forkel and Johann-Mattis List. 2020. CLDF-Bench: Give your cross-linguistic data a lift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6995–7002, Marseille, France. European Language Resources Association.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018b. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1):180205.

Hans Goebl. 2017. Dialectometry. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 123–142. Wiley.

Georg Hadzidakis. 1892. *Einleitung in die Neugriechische Grammatik*. Breitkopf & Härtel, Leipzig.

Harald Hammarström and Robert Forkel. 2022. Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information. *Semantic Web Journal*, 13(6):917–924.

Martin Haspelmath and Uri Tadmor. 2009a. *Loanwords in the world's Languages: a comparative handbook*. Mouton De Gruyter, Berlin.

Martin Haspelmath and Uri Tadmor, editors. 2009b. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at https://wold.clld.org, Accessed on 2026-02-26.

Wilbert Heeringa and Angelika Braun. 2003. The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances. *Computers and the Humanities*, 37(3):257–271.

Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In *Proceedings of the Workshop on Linguistic Distances*, pages 51–62, Sydney, Australia. Association for Computational Linguistics.

Wilbert Heeringa and Jelena Prokić. 2017. Computational Dialectology. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 330–347. Wiley.

Wilbert Heeringa, Vincent Van Heuven, and Hans Van de Velde. 2023. LED-A: Levenshtein Edit Distance App [computer program]. Available online at https://www.led-a.org/, Accessed on 2026-02-26.

Wilbert Jan Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance (Groningen Dissertations in Linguistics 46)*. Ph.D. thesis, University of Groningen, Groningen.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, Robert Tegethoff, Matilde Serangeli, Nikos Liosis, Krzysztof Stroński, Kim Schulte, Ganesh Kumar Gupta, Wolfgang Haak, Johannes Krause, Quentin D. Atkinson, Simon J. Greenhill, Denise Kühnert, and Russell D. Gray. 2023. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656):eabg0818.

Geoffrey C. Horrocks. 2010. *Greek: a history of the language and its speakers*, 2nd ed edition. Wiley-Blackwell, Oxford ; Malden, Mass. OCLC: ocn416717812.

ILIK. 2024. *Historical Dictionary of the Cappadocian Dialects [in Greek]*. Athens Academy, Athens. Available online at https://ilik.academyofathens.gr, Accessed on 2026-02-26.

ILNE. 2026. The Historical Lexicon of Modern Greek [in Greek]. Thesaurus Linguae Graecae Project, University of California, Irvine. Available online at https://stephanus.tlg.uci.edu/ilne, Accessed on 2026-02-26.

Gereon A. Kaiping, Melvin S. Steiger, and Natalia Chousou-Polydouri. 2022. Lexedata: A toolbox to edit CLDF lexicaldatasets. *Journal of Open Source Software*, 7(72):4140.

Georgia Katsouda. 2024. Dialect Lexicography: Greek Bibliography (II) [in Greek]. *Lexicographic Bulletin*, 27:355–379.

Georgia Th. Katsouda. 2021. *Stsyriana: A Synchronic Description and Analysis of the Linguistic Idiom of Skyros [in Greek]*. Stamoulis Publications, Athens.

Mary Ritchie Key and Bernard Comrie. 2023. The Intercontinental Dictionary Series. Max Planck Institute for Evolutionary Anthropology. Available online at https://ids.clld.org, Accessed on 2026-02-26.

Maxim L. Kisilier and Tanguy Solliec. in press. Is Azov Greek a variety of Pontic? Preliminary remarks. *Greek Around the World: Papers from the 9 February 2024 Workshop*. HAL Id: hal-04921554.

William A. Kretzschmar. 2017. Linguistic Atlases. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 57–72. Wiley.

Doris Kyriazis and Aristotelis Spyrou. 2011. The Greek Linguistic Idioms of Albania [in Greek]. *Modern Greek Dialectology*, 6:175–199.

Alfred Lameli. 2009. Linguistic atlases - traditional and modern. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space: Theories and Methods*, pages 567–592. Walter de Gruyter.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

Johann-Mattis List and Robert Forkel. 2022. Automated identification of borrowings in multilingual wordlists. *Open Research Europe*, 1:79.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

Johann Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2025. *CLLD Concepticon 3.4.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at https://concepticon.clld.org, Accessed on 2026-02-26.

Io Manolessou. 2019. The Historical Background of the Asia Minor Dialects. In Angela Ralli, editor, *The Morphology of Asia Minor Greek*, pages 20–65. Brill.

Io Manolessou, Stamatis Beis, and Christina Basea-Bezantakou. 2012. The phonetic transcription of the Modern Greek dialects [in Greek]. *Lexikografikon Deltion*, 26:161–222.

Dimitra Melissaropoulou, Stavros Bompolas, and Charalampos Tsimpouris. 2022. Digital cartography in the service of preservation of cultural linguistic heritage: Implementing the electronic dialectal atlas of Cappadocian Greek. *Scientific Culture*, 8(2):135–146.

Claudine Moulin. 2009. Dialect dictionaries - traditional and modern. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space: Theories and Methods*, pages 592–612. Walter de Gruyter.

John Nerbonne. 2009. Data-Driven Dialectology. *Language and Linguistics Compass*, 3(1):175–198.

John Nerbonne and Peter Kleiweg. 2007. Toward a dialectological yardstick*. *Journal of Quantitative Linguistics*, 14(2-3):148–166. Number: 2-3.

John Nerbonne, Jelena Prokić, and Martijn Wieling. 2021. Dialectology for Computational Linguists. In Marcos Zampieri and Preslav Nakov, editors, *Similar Languages, Varieties, and Dialects. A Computational Perspective*, 1 edition, pages 96–118. Cambridge University Press.

John Nerbonne and Martijn Wieling. 2017. Statistics for Aggregate Variationist Analyses. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 400–414. Wiley.

Nikolaos Pantelidis. 2001. Peloponnesian dialectal accent and Standard Modern Greek [in Greek]. In *Proceedings of the Fourth International Conference of Greek Linguistics*, pages 480–486, Thessaloniki. University Studio Press.

Anthimos Papadopoulos. 1926. *Grammar of the Northern Idioms of the Modern Greek Language [in Greek]*. P. D. Sakellariou Press, Athens.

Jelena Prokić and John Nerbonne. 2008. Recognising Groups among Dialects. *International Journal of Humanities and Arts Computing*, 2(1-2):153–172. Number: 1-2.

Stamatios B. Psaltis. 1905. *Thracian Studies or A Study of the Linguistic Idiom of the City of Saranta Ekklisies [in Greek]*. P. D. Sakellariou Press, Athens.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and Computational Cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Matthew Scarborough. 2019. Cognacy and Computational Cladistics: Issues in Determining Lexical Cognacy for Indo-European Cladistic Research. In *Dispersals and Diversification: Linguistic and Archaeological Perspectives on the Early Stages of Indo-European*, pages 179 – 208. Brill, Leiden, The Netherlands.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Russell Barlow, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Johannes Englisch, Angela Chira, Annemarie Verkerk, Russell Dinnage, Luke Maurits, Sam Passmore, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browen, Hoju Cha , Lennart Chevallier, Swintha Danielsen, Hugo De Vos, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Grace Ephraums, Giada Falcone, Anna-Maria Fehn, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Samuel Griggs, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Manuel Rüdisühli, Alexandra Marley,

Tânia R. A. Martins, Marvin Leonard Martiny, Celia Mata German, Suzanne Van Der Meer, Jacob Menschel, Jaime Montoya, Michael Müller, Saliha Muradoglu, HunterGatherer, David Nash, Kelsey Neely, Johanna Nickel, Miina Norvik, Olga Olina, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Leah Pappas, Jesse Peacock, India O.C. Pearey, Naomi Peck, Jana Peter, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Tihomir Rangelov, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Kristian Roncero Toledo, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Dineke Schokkin, Jeff Siegel, Jane Simpson, Amalia Skilton, Hilário De Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Daniel Wikalier Smith, Nicholas Williams, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye , Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Olena Shcherbakova, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank v1.0. Zenodo.

George Starostin. 2011. The Global Lexicostatistical Database. Higher School of Economics and Santa Fe Institute. Available online at https://starlingdb.org/new100/, Accessed on 2026-02-26.

Morris Swadesh. 1955. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21(2):121–137.

Manolis Triandaphyllidis. 1938. *Modern Greek Grammar. Volume A: Historical Introduction [in Greek]*. Manolis Triandaphyllidis Foundation, Thessaloniki.

Peter Trudgill. 2003. Modern Greek dialects: A preliminary classification. *Journal of Greek Linguistics*, 4(1):45–63.

Christos Tzitzilis. 2000. Modern Greek dialects and Modern Greek dialectology [in Greek]. In A.-Ph Christidis and others, editors, *The Greek Language and its Dialects [in Greek]*, pages 15–22. Ministry of National Education and Religious Affairs & Centre for the Greek Language, Athens.

Jacques Van Keymeulen. 2017. The Dialect Dictionary. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 39–56. Wiley.

Tessa Y. Vermeir, Marc Allassonnière-Tang, and Guillaume Segerer. 2024. LA80: A Lexical Database of 10 Bantu A80 Languages. *Journal of Open Humanities Data*, 10:42.

Søren Wichmann, Eric W. Holman, Cecil H. Brown, Matthew S. Dryer, and Qibin Ran. 2025. The ASJP Database (version 21). Available online at https://asjp.clld.org, Accessed on 2026-02-26.

Martijn Wieling and John Nerbonne. 2015. Advances in Dialectometry. *Annual Review of Linguistics*, 1(1):243–264.

A Speech Resource for the Pontic Greek Dialect: Transcription Choices and Baseline ASR Evaluation

Rodanna Konstantinidou[1], **Chara Tsoukala**[2], **Vivian Stamou**[2]
Voula Giouli[1,2], **Stella Markantonatou**[2]
[1]Aristotle University of Thessaloniki, [2]ILSP, Athena R.C.
{rkonstaa,pgiouli}@del.auth.gr, {chara.tsoukala,vivian.stamou,marks}@athenarc.gr

Abstract

Pontic Greek is a living but endangered Modern Greek dialect that lacks publicly available AI-oriented speech resources and ASR benchmarks. This work reports on the first systematic inference-only (zero-shot) ASR evaluation on authentic Pontic speech. Progress on Pontic ASR is hindered by two coupled challenges: the scarcity of transcribed speech data and the absence of a standardized orthography, which makes it difficult to create consistent reference transcriptions for evaluation. We address these challenges by releasing a new speech corpus of contemporary Pontic as spoken in Northern Greece, derived from natural conversations and provided with manual, utterance-level, time-aligned transcriptions. To reduce annotator bias and increase practical usability, we collect community evidence on written-form preferences via a small questionnaire and use the observed patterns to guide a consistent Greek-script transcription scheme. We use this corpus to perform inference-only (zero-shot) ASR evaluation, benchmarking four state-of-the-art pretrained speech recognition models under a unified evaluation protocol. Results show that zero-shot recognition remains challenging, establishing baseline figures and underscoring the need for dialect-specific data and adaptation.

Keywords: non-standard languages, low-resource ASR, Pontic dialect

1. Introduction

We initiate an effort to provide Pontic Greek, a living, endangered dialect of Modern Greek, with AI-oriented tools and resources to contribute to its vitality. In this work, the first attempt to evaluate Automatic Speech Recognition (ASR) tools against authentic Pontic spoken data is presented. The enterprise raises questions about the orthographic representation of the dialect; this study makes a first attempt to discuss this issue by consulting a small part of the community and it appears to reveal interesting requirements.

Our evaluation relies on ASR systems trained on the related well-resourced language variety, namely Standard Modern Greek (SMG). Using technology developed for a standard variety in order to work with related dialectal data is almost a rule-of-thumb since standard varieties tend to have more resources than the related dialects. However, the fact is that the more distant the dialect from the standard, the less accurate results are obtained. Pontic turns out to be an excellent example of this situation.

Our contributions are threefold: (i) a new Pontic Greek speech corpus[1] (Northern Greece) with utterance segmentation and time-aligned manual transcriptions (Praat/TextGrid), enabling reproducible speech-text pairing for ASR evaluation; (ii) community-informed, internally consistent transcription conventions in Greek script (favoring broadly usable forms and avoiding diacritics) to support reliable ASR references and other downstream work; and (iii) the first systematic inference-only (zero-shot) ASR baseline on authentic Pontic speech, benchmarking state-of-the-art pretrained speech recognition models using WER and CER.



Figure 1: Geographic extent of the Pontus region until 1922 (olive) and the main areas of Pontian settlement in Greece following migration (blue).

The paper is structured as follows: An introduction to the history of Pontic is given along with a brief description of its phonological features that differentiate it from SMG. Next, we discuss relevant work with a focus on work on other Modern Greek dialects. The methodology we used to collect, transcribe and prepare the data for the ASR experiments is presented next followed by a description of the experimental work. We conclude with a brief discussion of the results.

[1] https://huggingface.co/datasets/ilsp/pontic-speech-corpus

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 300–307
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

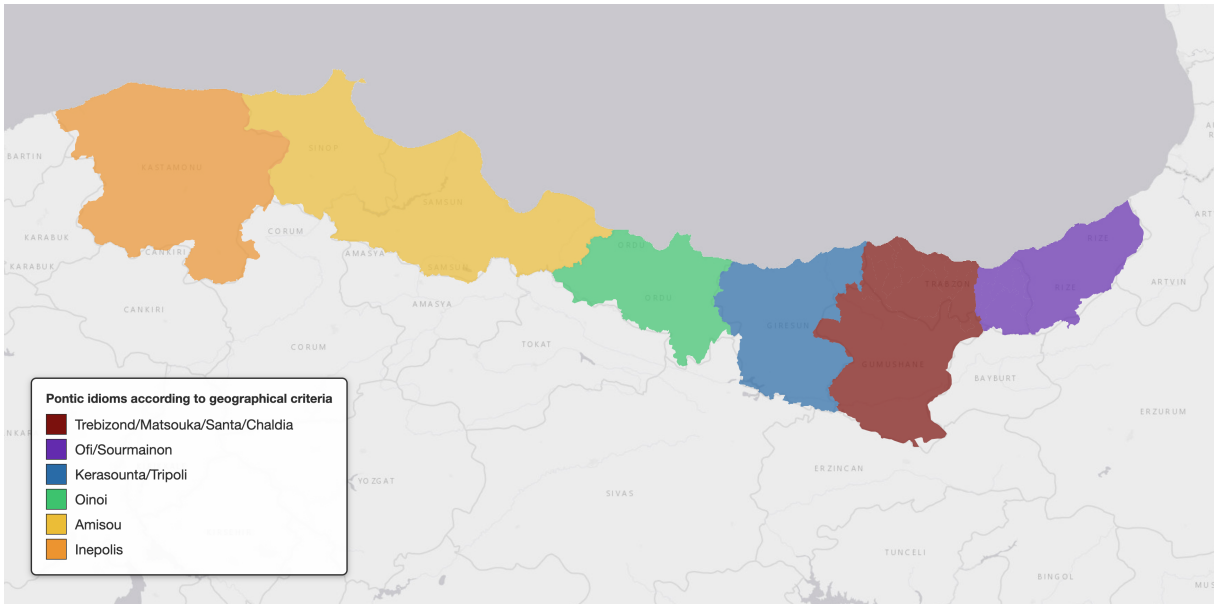


Figure 2: Pontic idioms according to geographical criteria (Papadopoulos, 1953)

2. About Pontic Greek

Pontic Greek (ISO 639-3: pnt; Glottolog: pont1253) is a Modern Greek dialect that belongs to the Indo-European language family. The dialect was spoken until 1922 along the southern coast of the Black Sea and in the interior of Asia Minor, extending from Inepolis (in the west) to Rizounta and Colchis (in the east) as seen in Figure 1. Beyond the coast, Pontic was spoken south of Trabzon, particularly in the regions of Gemoura, Matsouka, Santa, Kromni, Chaldia, and Cherianon, as well as in the mining region, the most prominent of which being Gümüş-Maden and Akdağmadeni (Tombaidis, 1988). In several areas, Pontic speakers (Pontians) coexisted with Turkish-speaking populations, a factor that significantly influenced the development of the dialect (Drettas, 1997; Dimela, 2013).

According to Papadopoulos (1953), the Pontic dialect was divided into six different idioms based on geographical criteria: (i) the idiom of Trebizond, Matsouka, Santa, and Chaldia (the most populous), (ii) Ofi and Sourmainon, (iii) Kerasounta and Tripoli, (iv) Oinoi, (v) Amisou, and (vi) Inepolis, as seen in Figure 2.

After the population exchange in 1923, following the signing of the Treaty of Lausanne, Pontian refugees settled in various areas of Greece, mainly in Northern Greece (Macedonia and Thrace) and, to a lesser extent, in Attica (Figure 1); Pontians also settled out of Greece, namely in the Caucasus region and Ukraine. Currently, the dialect is estimated to be spoken by approximately 400.000 speakers in Greece according to Ethnologue (Eberhard et al., 2026). UNESCO (UNESCO, 2010) classifies the Pontic dialect as “definitely endangered” as its use is declining and younger generations tend to adopt Standard Modern Greek (SMG).

In this study, we focus on the variety of Pontic as spoken today in Northern Greece, which is the most widespread contemporary form of the dialect.

A primary challenge for the development of ASR tools for the Pontic dialect is the total lack of publicly available datasets of recorded and transcribed speech. In addition to the lack of data, the internal variation of the dialect is significant.

Pontic Greek exhibits characteristic phonological features that complicate the development of ASR models since they differentiate it from SMG:
1. The vowels as /æ/ and /ø/, often represented orthographically with letters of the Greek alphabet plus diacritics (e.g. ä, ö); they originate from vowel coalescence phenomena (e.g. ια, εα > ä).
2. The consonants /ʃ/, /kʃ/, /pʃ/, typically rendered as ǒ, ξ̌, ψ̌, which are absent from SMG phonology.

Like many other dialects of Modern Greek, Pontic includes spoken features like the ones listed above, that are not represented with the orthographic conventions of SMG. Several proposals have been made regarding the orthography of Pontic: Papadopoulos (1955, 1958), Symeonidis and Tombaidis (1999), Oikonomidis (1958), Valavanis (1892), and Parcharidis (1885), in a comprehensive study (Katsouda, 2012). However, these proposals do not agree among themselves, and, most importantly, they have not been adopted by the Pontic community widely, see Section 4.2.1: Pontic has not been standardised so far.

Overall, Pontic Greek constitutes a representa-

tive case of a low-resource dialect, where the combination of limited data availability, phonological and orthographic variability, and lack of standardisation poses significant challenges for the development of reliable ASR systems.

In this paper, where we present the first systematic evaluation of ASR models for the Pontic Greek language, we pay particular attention to transcription choices by consulting the community in an effort to better capture the phonetic characteristics of the dialect and improve model compatibility.

3. Related Work

Recent progress in ASR technologies has been driven by large-scale multilingual models capable of handling multiple languages with minimal task-specific adaptation. Prominent examples include Whisper (Radford et al., 2022) and XLS-R (Babu et al., 2022), which have demonstrated strong performance on a wide range of languages, including low-resource settings. These models benefit from training on large and diverse speech corpora, enabling robust generalisation and competitive zero-shot capabilities.

Dialectal ASR turns out to be rather challenging. Dialects are typically under-resourced. While turning to ASR systems of the related well-resourced varieties -typically of the standard variety- is the obvious choice, recent research suggests that dialects behave as distinct acoustic-linguistic domains rather than minor variants of a canonical form (Torgbi et al., 2025; Dhasmana et al., 2026). Likewise, the MADASR initiative highlights persistent performance gaps across closely related South Asian varieties (Singh et al., 2023).

In the case of SMG, for which sufficient data is available, several studies report relatively strong ASR performance. Fine-tuned variants of Whisper achieve WER in the range of 13-16%, while XLS-R-based models achieve comparable or slightly better results on benchmark datasets such as Common Voice (Ardila et al., 2020). Recent work leveraging weakly supervised data, such as large-scale podcast corpora, demonstrates that increasing data volume and domain diversity can significantly improve ASR performance for SMG (Paraskevopoulos et al., 2024).

However, performance degrades substantially with dialectal distance from the Standard variety. Vakirtzian et al. (2024) provide the first systematic evaluation of ASR models on multiple Greek dialects, including Aivaliot, Cretan, Griko, and Messenian. Their results show that state-of-the-art models that achieve low WER on SMG (11-14%) experience severe degradation with dialectal data: WER often exceeds 50% and even surpasses 100% in zero-shot settings. In the same line, a recently published expanded benchmark for low-resource Modern Greek dialects (Tsoukala et al., 2026) covers Aperathiot, Cretan, Lesbian, and Cappadocian, which are varieties with varying degrees of divergence from SMG. The benchmark provides transcriptions following SMG-based orthographic conventions, while preserving dialectal lexical and morphophonological forms. Zero-shot results with several ASR models reveal a clear performance gradient with dialectal distance from SMG, with best WERs ranging from about 60-70% for southern dialects, which are close to SMG, to over 80% for Lesbian and nearly 97% for Cappadocian. Fine-tuning substantially reduces error rates (up to 47% relative WER improvement), with Cappadocian remaining the most challenging variety (best WER 68.17%).

Despite the growing body of work on Greek dialects, Pontic Greek remains largely underexplored in the context of ASR, and there is currently no systematic evaluation of speech recognition models specifically targeting this variety. Nevertheless, evidence from closely related NLP tasks suggests that Pontic Greek poses significant challenges for modern language technologies. In particular, (Chatzikyriakidis et al., 2025) demonstrates that large language models (Llama-3-8B (Grattafiori et al., 2024), Llama-3.1-8B, Krikri-8B (Roussis et al., 2025)) exhibit substantially degraded performance on dialectal Greek, with Pontic Greek being among the most difficult cases. The authors further show that targeted fine-tuning on dialectal data leads to notable improvements, highlighting both the limitations of pretrained models and the importance of dialect-specific adaptation. Given the shared challenges across speech and text modalities, such as lexical variation, non-standard orthography, and scarcity of annotated speech data, these findings provide strong indirect evidence that ASR systems are also likely to struggle with Pontic Greek. This gap in the literature underscores the need for dedicated evaluation and modeling efforts for this dialect.

4. Methodology

The recordings were carefully collected, transcribed, and annotated to create a consistent and reliable dataset for ASR evaluation. Manual transcription ensured orthographic standardisation and precise time-aligned segmentation of the audio. The resulting dataset was then used to evaluate ASR performance using the Whisper and XLS-R models, with WER and CER as evaluation metrics. The following subsections describe the decisions and procedures employed to ensure accuracy, consistency, and fidelity to the Pontic Greek dialect.

4.1. Data Collection

Three native speakers of the Pontic dialect participated in the audio recordings, namely two females and one male, aged between 70 and 80. Originally from Neokaisareia (currently known as Niksar), the speakers were raised in Kallithea and Lofos, two villages in the Elassona region, and share a common variety of the Pontic dialect. This variety reflects the spoken form currently spoken by Pontic communities in different regions of Greece, mainly in Northern Greece, and does not correspond to historical or regional variants such as Ofitic.

The recordings were designed to capture spontaneous, natural speech through informal discussion. Participants were prompted to speak about topics related to everyday life, cultural memories, traditional food, local customs, games, and schooling. The aim was to elicit authentic dialectal usage in a conversational context, producing speech that reflects the natural patterns, vocabulary, and pronunciation of the Pontic dialect as it is spoken today.

The total recording duration is 1 hour, 26 minutes, and 38 seconds. It was conducted in Katerini, Greece, where the speakers currently reside, in a quiet indoor environment with minimal background noise to ensure high audio quality. This controlled setting allowed for clear capture of speech, facilitating subsequent transcription, segmentation, and annotation for the creation of a reliable and representative dataset for ASR evaluation.

4.2. Transcription & Annotation

This section describes the methods used to transcribe, segment, and annotate the recordings, as well as how the evaluation of models on Pontic Greek was conducted. It provides details on the orthographic decisions, manual transcription process, and time-aligned segmentation that together produced a consistent and reliable dataset for ASR evaluation.

4.2.1. Orthographic Decisions & Community Validation

Due to the lack of a standardized orthography, written representation practices vary considerably among speakers of Pontic, rendering the establishment of consistent and reliable reference transcriptions for ASR evaluation challenging. Our methodology prioritised developing a consistent set of transcriptions while minimising annotator bias in orthographic choices.

To inform the transcription strategy, we designed a written form questionnaire that targeted native speakers of the Pontic dialect. Ten participants aged between 40 and 80, with varying levels of daily dialect use (3/10 speakers speak the dialect daily, 4/10 occasionally, and 3/10 rarely), contributed written responses. The questionnaire included isolated lexical items, focusing on key phonological phenomena that would pose a problem in transcription. Participants were instructed to write words as they would naturally represent them, without looking them up in dictionaries or using copy and paste; the aim was to obtain evidence that captures authentic variation in orthographic practices.

The analysis of the responses identified recurring spelling patterns and informed the orthographic strategy for the transcription of the spoken data. The consonants "/ʃ/", "/pʃ/", "/kʃ/", and "/tʃ/", and the vowel "/æ/" at the ends of words, where the phenomenon of synizesis is observed are not represented in the Greek alphabet. Their representation in the proposed Pontic orthographies is done with Greek letters plus diacritics. Table 1 shows these representations and the responses of the speakers of Pontic. Regarding the phoneme /ʃ/ (e.g. χέρι=hand), the responses of speakers of Pontic were divided between renditions with "σ" and "χ", with variations such as "σσιέρ", "χεριν", and "σshέρι". Although there was no clear majority among the speakers, most researchers agree on the use of "χ" (Katsouda, 2012). For this reason, the spelling "χ" was chosen to ensure consistency with theoretical guidance and maintain scientific validity. /pʃ/ (e.g. ψυχή=soul) and /tʃ/ (e.g. καλατσεύω=speak) were represented by the majority of speakers with "ψ" and "τσ", with minor variations such as "ψι", "ψύ", "ψυη", or "καλατζευω" and "καλατshεύω". Regarding /kʃ/ (e.g. ρίχνω=throw), speaker responses varied, partly due to influence from SMG which uses the verb forms ρίχναμε, ερίχναμε; as a result, only five interesting spelling choices were obtained. Among these, the majority favored "ξ" over variants such as "έκχυναμε" and "εκσσσιναμε". Finally, although the vowel /æ/ (e.g. ποτήρια=glasses) cannot be represented with the Greek alphabet, speaker responses showed strong agreement in rendering it as "ε".

These orthographic decisions favored the most frequently attested forms, adopted simplified representations, which are all characters of the Greek alphabet, and eliminated the use of diacritics. This approach provided an internally consistent reference set suitable for evaluating ASR models while reflecting the authentic linguistic patterns of Pontic Greek.

4.2.2. Transcription, Segmentation and Annotation

The transcription of the Pontic Greek recordings was performed semi-automatically, combining automatic transcription with manual post-editing to ensure accuracy and consistent application of the orthographic conventions. Whisper large-v2 was

IPA	Prompt	Responses (10)	Decision	Papadopoulos 1955/58	Parcharidis 1885	Valavanis 1892	Oikonomidis 1908	Oikonomidis 1958	Symeonidis–Tombaidis 1999
/ʃ/	e.g. χέρι (hand)	σιέρ (2), σεερ (1), χ̌έρ΄ (1), χερα (1), σσσερ (1), σshέρι (1), χεριν (1), χέρι (1), σσερ (1)	χ	χ̌	χ̇	χ	χ̇	χ̂	χ̌
/pʃ/	e.g. ψυχή (soul)	ψυ (2), πσι (1), ψή (1), πσσσσυ (1), ψυη (1), ψύ (1), ψshιν (1), ψυμ (1), ψυν (1)	ψ	ψ̌	ψ̇	ψ′	ψ̇, ψ′χ̇	ψ̂, ψ′χ̂	ψ̌
/tʃ/	e.g. καλατσεύω (speak)	καλατσευω (3), καλατσέυω (2), λαλώ (1), καλατσσσεύω (1), καλατσέβω (1), καλατσhεύω (1), καλατζευω (1)	τσ	τσ̆	τ̇, κ̇	τσ̇, τ́σ	τσ̇	τσ̂	τσ̆
/kʃ/	e.g. ρίχνω (throw)	έκχ̆υναμε (1), εκσσσιναμε (1), έξιναμ (1), έξυναμε (1), εξύγαμεν (1)	ξ	ξ̆	ξ̣	ξ′	ξ̇, ξ′χ̇	ξ̂	ξ̆
/ae/	e.g. ποτήρια (glasses)	ποτήρια (2), ποτήρε (2), κενό (2), ποτηρε (1), ποτυρε (1), ποτήρεα (1), ποτίρε (1)	ε	α̤	ᵉα	ια	ä	ä	α̤

Table 1: Comparison of consonant and vowel representations across Pontic Greek orthographic systems. (Katsouda, 2012)

used to generate rough transcriptions as an initialisation step. Automatically-generated transcriptions were carefully reviewed and corrected by a linguist who is a speaker of the Pontic dialect, ensuring adherence to the spoken language while following the orthographic guidelines established through the community validation step.

Audio segmentation divided the recordings into individual utterances with the Praat tool (Boersma and Weenink, 2022). Each utterance was aligned with its corresponding audio segment to provide accurate timing information. The transcriptions and their time-aligned segments were stored in `TextGrid` files, which labeled each utterance with start and end times. From the original 1 hour, 26 minutes, and 38 seconds of recording time 39 minutes and 58 seconds of speech free from significant background noise were obtained.

The combination of manual verification, Whisper-assisted initialisation, and precise segmentation resulted in a high-quality dataset suitable for evaluating ASR models while maintaining linguistic and orthographic accuracy.

4.3. Corpus Construction and ASR Evaluation

The manual transcription and time-alignment process described above resulted in a set of Praat `TextGrid` files, where each utterance is associated with start/end timestamps and an orthographic transcription in the Greek script. To enable ASR benchmarking, we converted these annotations into a standard speech-corpus format consisting of audio–text pairs. The resulting corpus is thus an utterance-segmented version of the original recording, aligned to the dialectal transcriptions.

Preprocessing. To ensure consistency with common ASR benchmarking practices for Greek dialectal data, we applied lightweight preprocessing similar to prior work on Modern Greek dialect ASR evaluation resources (Vakirtzian et al., 2024)[2]. Audio was converted to mono wav at 16 kHz. On the text side, we lowercased and removed punctuation. Note that we evaluate dialectal transcription rather than dialect-to-standard normalisation.

Segmentation and filtering. Utterance boundaries were defined in Praat and exported from the TextGrid annotations. During this step, non-speech material (e.g., long pauses and other non-linguistic intervals) was excluded. The resulting segments constitute the evaluation units used for all ASR runs.

Models and inference-only setup. Given the small size of the corpus (single recording session, limited speaker and domain diversity), we report inference-only (zero-shot) performance without dialect-specific adaptation, to avoid overfitting and to quantify the out-of-the-box robustness of current ASR models on Pontic Greek. We evaluated two families of pretrained ASR models without dialect-specific adaptation: (i) Whisper-based (Radford et al., 2022) multilingual sequence-to-sequence models (Large-v2, Large-v3, Medium), and (ii) a Greek-adapted XLS-R (Babu et al., 2022) model[3] (wav2vec2/CTC-style).

All models were run on the utterance-segmented corpus created from the `TextGrid` annotations,

[2] https://github.com/athena-ilsp/greek-dialects-asr

[3] huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-greek

producing a hypothesis transcription per segment. For Whisper, we used standard transcription inference with the language set to Greek (language="el") and no user-supplied initial prompt. No external language model or additional decoding constraints were used. For evaluation, both references and hypotheses were lowercased and stripped of punctuation; no further text normalization was applied. We computed Word Error Rate (WER) and Character Error Rate (CER) on the normalised text (lowercased, punctuation removed). Reporting both WER and CER is particularly useful for dialects, where orthographic variability and dialect-specific phonological realisations may disproportionately impact word-level errors, while character-level scoring can better reflect partial matches and subword robustness.

Model	WER (%)	CER (%)
Whisper large-v2	91.24	52.77
Whisper large-v3	**86.59**	**47.72**
Whisper medium	105.47	65.41
XLS-R-53-greek	111.70	98.88

Table 2: Inference-only ASR results on the Pontic Greek corpus.

5. Experimental Results

Table 2 reports inference-only ASR performance on the Pontic Greek corpus. Overall, zero-shot recognition is challenging: WERs are high across all models, and in some cases exceed 100%, indicating that the number of edit operations (substitutions, insertions, deletions) surpasses the number of reference words. These findings are consistent with prior work on Greek dialectal ASR evaluations (Vakirtzian et al., 2024; Tsoukala et al., 2026), where zero-shot evaluations on Cappadocian and Griko similarly yielded WER values above 100%, underscoring the broader difficulty of recognizing non-standard varieties without dialect-specific adaptation.

The inference-only results indicate that the Whisper-based models outperform the Greek-adapted XLS-R model on the Pontic Greek corpus. Among the Whisper models, large-v3 achieves the best performance, with a WER of 86.59% and a CER of 47.72%, followed by large-v2. The medium-sized Whisper model exhibits substantially higher error rates, suggesting that model capacity strongly impacts zero-shot recognition quality for this low-resource dialect.

The Greek-adapted XLS-R model performs poorly in this zero-shot setting (WER 111.70%, CER 98.88%), indicating limited generalization to Pontic speech despite adaptation to SMG. The very high CER suggests that the model frequently fails to recover even approximate character sequences under the combined acoustic, lexical, and orthographic mismatch induced by the Pontic dialect.

Taken together, these results establish strong zero-shot baselines and indicate that Pontic Greek is difficult for current general-purpose ASR. By releasing an utterance-segmented and time-aligned Pontic speech corpus with consistent transcriptions, we aim to support reproducible benchmarking and to enable future work on dialect-specific adaptation and evaluation protocols.

6. Discussion

It was already known that Pontic Greek is dialectologically distant from SMG as compared to other Modern Greek dialects. This fact is demonstrated in the ASR inference-only results, for instance, Pontic results match those of Cappadocian, another Modern Greek dialect of Asia Minor that is quite distant from SMG dialectologically (Tsoukala et al., 2026). These results underline the need for more Pontic data as well as the application of a variety of techniques to develop ASR models for the Pontic dialect.

Another interesting point touched upon in this work is the orthography issue. Despite the small number of native speakers who responded to a first simple questionnaire about the orthography of the dialect, the tendency that emerged is clear: native speakers of unstandardised varieties do not adhere to detailed, more precise orthographies and tend to cling to the orthography of the Standard variety which is taught in school and has all the necessary technical support (of course, dialectal speakers speak both varieties, Pontic and the Standard). This is a tendency that should be investigated in more depth and, probably, taken seriously into account in the development of AI-oriented technology for Pontic.

7. Conclusion

We have presented the first evaluation of ASR technology on data from Pontic Greek, which is a living but endangered dialect of Modern Greek. Although Pontic Greek is spoken by a community of about 400K people, it is considered endangered as transmission to newer generations has largely ceased. Our results show that substantial data requirements and the adoption of techniques developed for low-resource languages are necessary to achieve high-quality ASR performance for Pontic.

Nevertheless, this study is the first of its kind and establishes an initial benchmark for the development and evaluation of ASR technologies that

may contribute to the vitality of this Modern Greek dialect.

Future work is planned towards expanding the dataset by incorporating the speech of a larger number of native speakers of the dialect. In parallel, we plan to conduct a large-scale survey on orthographic preferences and to experimentally evaluate the effects of different orthographic conventions on transcription consistency and downstream tasks.

8. Limitations

We are aware that the body of the data is small, which has dictated inference-only experiments. In addition, the number of speakers contacted for othography issues is small as well as the range of phenomena discussed.

9. Ethical considerations

Speech data were collected with the speakers' written permission for recording and public release. The corpus is openly available under a Creative Commons license.

10. Acknowledgements

The work presented in this paper was partially supported by the Aristotle University of Thessaloniki, Greece (Grant ELKE-AUTh-13337). We would like to thank the anonymous reviewers for their valuable comments and constructive feedback.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech*.

Paul Boersma and David Weenink. 2022. *Praat: doing phonetics by computer*. Version 6.2.06, retrieved 23 January 2022 from http://www.praat.org/.

Stergios Chatzikyriakidis, Dimitris Papadakis, Sevasti-Ioanna Papaioannou, and Erofili Psaltaki. 2025. Grdd+: An extended greek dialectal dataset with cross-architecture fine-tuning evaluation.

Akriti Dhasmana, Aarohi Srivastava, and David Chiang. 2026. Dialect matters: Cross-lingual asr transfer for low-resource indic language varieties.

Eleonora Dimela. 2013. Ta tourkika epithetika daneia stis dialektous tis mikras asias: mia proti prosegnisi. *Patras Working Papers in Linguistics*, 3:92–109.

Georges Drettas. 1997. *Aspects pontiques*. Association de recherches pluridisciplinaires, Paris.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2026. Ethnologue: Languages of the world.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models.

Georgia Katsouda. 2012. Dialectal lexicography: review and issues. *Leksikografikon Deltion*, 26:133–159.

Dimosthenis Oikonomidis. 1958. *Grammatiki tis Ellinikis dialektou tou Pontou*. Academy of Athens, Athens.

Anthimos Papadopoulos. 1953. *Characteristika tis Pontiakis dialektou*, volume 18. Epitropi Pontiakokn Meleton.

Anthimos Papadopoulos. 1955. *Historiki grammatiki tis Pontikis dialektou*. Committee of Pontic Studies [Archeion Pontou, Supplement 1], Athens.

Anthimos Papadopoulos. 1958. *Historikon Lexikon tis Pontikis dialektou*. Committee of Pontic Studies [Archeion Pontou, Supplement 3], Athens.

Georgios Paraskevopoulos, Chara Tsoukala, Athanasios Katsamanis, and Vassilis Katsouros. 2024. The greek podcast corpus: Competitive speech models for low-resourced languages with weakly supervised data.

Ioannis Parcharidis. 1885. Glossologika. ta en chrisei stoicheia. *Astir tou Pontou*, 1:27–28.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arXiv:2212.04356.

Dimitris Roussis, Leon Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2025. Krikri: Advancing open large language models for greek.

Abhayjeet Singh, Arjun Singh Mehta, Ashish Khuraishi K S, Deekshitha G, Gauri Date, Jai Nanavati, Jesuraja Bandekar, Karnalius Basumatary, Karthika P, Sandhya Badiger, Sathvik Udupa, Saurabh Kumar, Savitha, Prasanta Kumar Ghosh, Prashanthi V, Priyanka Pai, Raoul Nanavati, Rohan Saxena, Sai Praneeth Reddy Mora, and Srinivasa Raghavan. 2023. Model adaptation for asr in low-resource indian languages. *arXiv preprint arXiv:2307.07948*.

Charalambos Symeonidis and Dimitris Tombaidis. 1999. *I simerini dialektos tis Oukranias (periochis Marioupolis)*. Committee of Pontic Studies [Archeion Pontou, Supplement 20], Athens.

Dimitris Tombaidis. 1988. *I pontiaki dialektos*. Parartima periodikou "Archeion Pontou". Epitropi Pontiakon Meleton.

Melissa Torgbi, Andrew Clayman, Jordan J. Speight, and Harish Tayyar Madabushi. 2025. Adapting whisper for regional dialects: Enhancing public services for vulnerable populations in the United Kingdom. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 29–38, Abu Dhabi, UAE. Association for Computational Linguistics.

Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos, and Stella Markantonatou. 2026. Extending asr evaluation resources for modern greek dialects. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects, Rabat, Morocco. Association for Computational Linguistics.*

UNESCO. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech recognition for greek dialects: A challenging benchmark. In *Proc. Interspeech 2024*.

Ioannis Valavanis. 1892. *Zonta mnimeia tis ana ton Ponton idiotikis. Tmema A', Lexilogion*. Typographeion Perri, Athens.

First Steps in ASR for Cypriot Greek: Challenges and Insights

Vivian Stamou[1], **Spyros Armostis**[3], **Antigoni Klimi**[2]
Georgios Paraskevopoulos[2], **Vassilis Katsouros**[2] , **Antonios Anastasopoulos**[1,4]
[1]Archimedes/Athena R.C., Greece
[2]Institute for Language and Speech Processing/Athena R.C., Greece
[3]University of Cyprus, Cyprus
[4]George Mason University, USA
{vistamou, antigoni.klimi, g.paraskevopoulos, vsk}@athenarc.gr
armosti.spyros@ucy.ac.cy, antonis@gmu.edu

Abstract

This paper presents the first automatic speech recognition (ASR) system for Cypriot Greek, a non-standardized variety of Modern Greek with distinctive lexicogrammatical and graphematic characteristics. We adapt Whisper, a state-of-the-art multilingual ASR model, to Cypriot Greek through fine-tuning on the Mozilla Common Voice spontaneous speech dataset for Cypriot Greek. The phonological and lexical divergence between Cypriot Greek and Standard Modern Greek poses significant challenges for mainstream ASR, particularly under conditions of limited training data and dialectal variation. Results demonstrate that whisper-medium achieved a best word error rate (WER) of 37.85%, while whisper-large-v3 consistently outperformed it, reaching a minimum WER of 33.93%. In the light of these findings, increased model size, combined with targeted fine-tuning on normalized dialectical data, significantly improves recognition accuracy, indicating that careful handling of orthographic and dialectical variation provides an effective path for ASR adaptation to low-resource varieties.

Keywords: low resource ASR, Cypriot Greek, dialectal resources

1. Introduction

State-of-the-art language and speech technologies are strongly data-driven, and language varieties with larger quantities of available data typically receive priority (Chin et al., 2023; Hedderich et al., 2020; Joshi et al., 2020). As a result, the speakers of non-dominant dialects have access to less technological support than the speakers of the dominant variety (Blasi et al., 2022). Therefore, promoting technological development for low-resource languages is vital for supporting language inclusion and preservation while enabling communities of minority languages to participate fully in the digital world (Mohanty et al., 2024).

In this work, we present the first ASR model for Cypriot Greek, a Modern Greek dialect with high vitality. In Section 2 Cypriot Greek is introduced. The adopted methodology is presented in Section 3, while Sections 4 and 5 report the experimental setup and corresponding results, respectively. Section 6 provides a brief overview of related work, and Section 7 concludes the paper.

2. Cypriot Greek

Cypriot Greek (CG) dialect belongs to the South-Eastern Greek dialect group. There are more than 700,000 speakers of CG in Cyprus and Cypriot communities abroad (UK, USA, Australia, Canada and South Africa). CG is considered the non standard variety of Modern Greek with the highest vitality (Statistical Service of Cyprus, 2011). The long time of isolation from other Greek-speaking areas led to substantial differences between Cypriot and Standard Modern Greek (SMG), occasionally rendering the two mutually unintelligible due to a host of phonological, morphological, lexical and syntactic differences. Importantly, this divergence also extends to the phonetic realization of segments and their coarticulation patterns, as shown by (Themistocleous, 2019), who successfully distinguish CG from SMG based on acoustic features of sonorant consonants and coarticulatory cues from adjacent vowels, achieving up to 81% classification accuracy using deep neural networks.

The official languages of the Republic of Cyprus are Greek and Turkish (the latter introduced to the island with the Ottoman conquest in 1571). Cypriot Greek is used in oral communication, while SMG functions in formal contexts and is the instructional language in public primary and secondary education. So far, CG has no standardized writing system, while attempts to establish language policies have never been further formalized (Papapavlou, 2010). Crucially, the SMG alphabet does not represent the distinctive sounds of the dialect, prompting several proposals for an appropriate writing system (Armostis et al., 2014). Nonetheless, there exists a body of medieval literature (13th–16th centuries), such as ‘Leontios Machairas: Recital about the sweet land of Cyprus entitled “Chronicle” (Syme-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 308–314
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

onidis, 2006) that represent regional linguistic features and sounds not present in SMG, such as postalveolar fricatives and affricates, and geminates. It is the case that CG writers often devise their own spelling conventions and preferences (e.g., different written renditions of [ˈʃ ɛ r i n] 'hand' as <σχέριν> or <σέριν>; Themistocleous, 2010).

Before 1974, Cypriot Greek displayed considerable internal variation (18 regional varieties according to Newton 1972) along a continuum that ranged from rural local dialects (basilects) to the urban variety (acrolect) (Newton, 1972; Contossopoulos, 1969). This continuum underwent homogenization due to demographic shifts after the Turkish military occupation and greater exposure to SMG (Tsiplakou et al., 2006). Nowadays, CG is largely uniform, with rare regional variants rapidly disappearing at the phonological, morphosyntactic, and lexical level (Tsiplakou, 2014).

3. Methodology

In this section, we describe the methodology employed to adapt ASR models to Cypriot Greek. We first present the data preparation pipeline, which involved manual transcription and input normalization. We then outline the fine-tuning configurations used to adapt Whisper and XLSR to the target variety.

3.1. Data collection

For ASR training, we employ version 23.0 of the Mozilla Common Voice Spontaneous Speech dataset for Cypriot Greek ('el-CY'). This resource comprises 1,284 audio clips, corresponding to approximately 11 hours of recorded speech produced by 10 speakers, transcribed and validated. The dataset is publicly accessible via the Mozilla Data Collective[1], with transcription statistics provided in Table 1.

Metric	Value
Prompts	146
Duration	39,313,800 [ms]
Avg. Transcription Length	350
Avg. Duration	30.62 [s]
Valid Duration	38,861.24 [s]
Total Hours	10.92 [h]
Valid Hours	10.79 [h]

Table 1: Transcription statistics of the el-CY dataset.

Specifically, Cypriot Greek audio recordings were transcribed by eight native speakers of the dialect. As it is a non-standardized variety, it has no codified orthography (Armostis et al., 2014). As discussed in Section 2, however, there exists a long-standing tradition in dialectal lexicography and edited literary works of representing Cypriot Greek in writing using the Greek script, supplemented with diacritics on top of certain letters to denote consonants that do not occur in SMG, as illustrated in Table 2[2]. The transcriptions were graphematically homogenized by two linguists to reduce spelling variation in the corpus; this was based both on existing dictionaries (Yiangoullis, 2014) and on the spelling conventions outlined in Armostis (2022).

Furthermore, given the absence of a standardized orthography for Cypriot Greek and the substantial variability observed in spontaneous speech transcripts, we applied a deterministic, rule-based text normalization pipeline prior to model training. The objective of this normalization was to reduce transcriptional variation that is orthographically irrelevant for ASR performance, while explicitly preserving dialectal identity and linguistically meaningful contrasts.

The normalization pipeline operates entirely at the Unicode text level and follows a fixed sequence of transformations to ensure reproducibility across training and evaluation splits.

1. **Unicode normalization and casing:** All transcripts are lowercased and normalized to Unicode NFC (Normalization Form C) form. Unicode NFD (Normalization Form D) decomposition is used internally to enable precise handling of combining diacritics, ensuring a single canonical representation for visually identical characters.

2. **Punctuation and whitespace normalization:** Quotation marks and apostrophes are unified into canonical forms, while all dash-like symbols are removed. Excess whitespace is collapsed into single spaces and trimmed (e.g., *"εν-τάξει"* → *"εντάξει"*).

3. **Removal of non-linguistic artefacts:** Non-linguistic elements such as noise markers, parenthetical comments, URLs, email addresses, phone numbers, and emojis are removed. Paralinguistic expressions are mapped to explicit placeholders (e.g., *"χαχαχα"* → *"(γέλιο)"*) to ensure consistent acoustic alignment.

4. **Normalization of elongations and repetitions:** Orthographic elongations are collapsed

[1] https://datacollective.mozillafoundation.org/datasets/cmflnuzz3xq2ok28b4o1ojacv

[2] While the transcription follows the conventions shown in Table 2, a different normalization strategy was used for the ASR experiments; further investigation into the impact of these distinct orthographic representations is reserved for future work.

Grapheme(s) & Case			Phonetic	Example	Comment
Lower	*Proper*	*Upper*	**Value**		
⟨ğ⟩	⟨Ğ⟩	—	[ʃ]	«ğέριν», «ίğια», «ğαğάρω»	
⟨ξ⟩	—	⟨Ξ̌⟩	[ʃ]	«ντούξ», «γιαβάξ»	word-finally
⟨ğğ⟩	⟨Ğğ⟩	⟨ĞĞ⟩	[ʃː]	«ğğίζω», «ğğεπάζω», «μελιğğιά», «μάğğαλλα»	
⟨ž⟩	⟨Ž⟩	—	[ʒː]	«πουžιάζω», «αμπαžούρ», «κκεράžια»	
⟨x̌⟩	⟨X̌⟩	—	[kʃ]	«x̌ιουρίζουμαι», «ταx̌ιά», «μοναx̌ιά», «γιαx̌ής»	
⟨ψ̌⟩	⟨Ψ̌⟩	—	[pʃ]	«ανιψ̌ιός», «κλεψ̌ιά», «καλουψ̌ής»	
⟨τž⟩	⟨Tž⟩	⟨TŽ⟩	[t͡ʃ]	«τžερίν», «χότžας», «ευτžή», «αυλατžιά»	
⟨ντž⟩	⟨Nτž⟩	⟨NTŽ⟩	[nd͡ʒ]	«ντžίζω», «φεντžάνιν», «καντžέλλιν», «πουντžιάζω»	
⟨τğ⟩	⟨Tğ⟩	⟨TĞ⟩	[tʃʰː]	«τğέκκιν», «Λατğιά», «κουτğιά»	
⟨τx̌⟩	—	⟨TX̌⟩	[tʃ]	«σάντουιτx̌», «ματx̌», «κλατx̌»	word-finally
⟨ȯ⟩	—	⟨Ȯ⟩	[o]	«πȯν», «μ̆δωκεν», «πȯννά»	from /u/ + /e/ coalescence

Geminates, i.e., long consonants, are indicated in spelling with a double letter: e.g., [m] is spelled as ⟨μμ⟩, as in ⟨σήμμερα⟩ [ˈsimːera] 'today'.

Table 2: Graphemes not present in the alphabet of Standard Modern Greek.

while preserving meaningful doubles (e.g., *"εεεεν"* → *"εν"*, *"ααα"* → *"αα"*). Word-initial duplicates caused by emphasis are simplified (e.g., *"εεεεσύ"* → *"εσύ"*).

5. **Polytonic-to-monotonic Greek conversion:** Legacy polytonic spellings are normalized by removing obsolete diacritics while preserving tonos and diaeresis (e.g., *"μητσŏί"* → *"μητσοί"*). Ill-formed spacing accent characters are either reattached to the appropriate vowel or discarded.

6. **Mixed Greek–Latin character correction:** Mixed Greek-Latin spellings commonly observed in Cypriot transcripts are corrected via targeted mappings from Latin to Greek characters, reducing orthographic fragmentation without altering pronunciation (e.g., *"τx̌ια"* or *"τzια"* → *"τζια"*).

7. **Dialect-preserving lexical canonicalization:** Cypriot-specific variants are mapped to canonical Cypriot forms rather than Standard Modern Greek (e.g., *"τζαι"*, *"τζε"* → *"τζ ̆αι"*). Optional normalization of legal geminates is applied depending on the experimental configuration.

To quantify the impact of normalization, we compute WER and CER between original and normalized transcripts, obtaining 15.8% and 3.8%, respectively. This indicates minimal, primarily orthographic modifications.

3.2. Data segmentation

To prepare the Cypriot Greek speech corpus for ASR fine-tuning, we segmented the audio into sentence-level units aligned with the corresponding transcriptions. Forced alignment was performed using the open-source *ctc-forced-aligner* toolkit[3] with the multilingual *mms-300m-1130* alignment model[4]. For each recording, the audio waveform was divided into overlapping windows (30 s with 2 s context) to generate CTC (Connectionist Temporal Classification) emissions, which were then aligned to the normalized transcript without requiring a pronunciation lexicon. The aligner produced token-level spans with start and end timestamps, from which sentence boundaries were derived.

The script outputs a Kaldi-style segmented dataset consisting of `wav.scp`, `text`, `utt2spk`, and `segments` files. In addition, a quality metric (`utt2score`) was computed for each utterance, combining alignment coverage, average log-probability, and duration heuristics. Segments with zero-length spans were automatically discarded, while utterances longer than 30 s were removed to match Whisper's input constraints. This procedure resulted in a clean set of sentence-level utterances with accurate time alignment, substantially improv-

[3] https://github.com/MahmoudAshraf97/ctc-forced-aligner

[4] https://huggingface.co/MahmoudAshraf/mms-300m-1130-forced-aligner

ing dataset quality and making it suitable for robust fine-tuning.

3.3. Model training

To assess the performance of state-of-the-art ASR models on Cypriot Greek, we selected two prominent ASR systems: XLSR-53 (Babu et al., 2021) and Whisper (Radford et al., 2023). XLSR-53, one of the first large multilingual wav2vec 2.0 models, was pretrained on 56,000 hours of speech across 53 languages. Whisper was trained on a much larger corpus, with the `small` model trained on 680,000 hours of weakly supervised multilingual data. For XLSR-53, we used a version further fine-tuned on Greek to improve performance on the target variety. Despite these pretraining efforts, both models exhibit reduced performance on dialectal varieties, particularly Cypriot Greek, whose phonological and lexical differences from SMG poses challenges for mainstream ASR.

To address this gap, we fine-tuned four Whisper variants on the Cypriot Greek dataset: `whisper-small`[5], `whisper-medium`[6], `whisper-large-v3`[7], and `whisper-large-v3-turbo`[8], as well as the `XLS-R-greek` model[9].

4. Experimental Setup

All experiments were conducted using the Hugging Face Transformers library[10]. Fine-tuning was performed on both the original and normalized Cypriot Greek corpus, which was split into training and evaluation subsets using an 80/20 ratio to ensure a representative validation set. We note that the train–test split is not speaker-independent.

Training proceeded for 2,048 update steps with a learning rate of 5×10^{-6}, a maximum input duration of 30,s, and an output length capped at 225 tokens. An effective batch size of 64 was used; the encoder was kept frozen while the decoder was fully trainable. To optimize GPU memory usage, gradient checkpointing and mixed-precision training (fp16) were enabled.

5. Results

In this section, we present the performance of the fine-tuned Whisper models on the Cypriot Greek test set, using Word Error Rate (WER) and Character Error Rate (CER) as evaluation metrics. Tables 3 and 4 summarize the WER and CER before and after fine-tuning for the *whisper-small*, *whisper-medium*, and *whisper-large-v3* variants, as well as the fine-tuned XLS-R-greek model, evaluated on both normalized and original transcripts.

Model	WER (%)		CER (%)	
	Before	After	Before	After
Whisper-small	81.27	**52.38**	49.21	**32.33**
Whisper-medium	69.03	**37.85**	39.69	**26.71**
Whisper-large-v3	58.33	**33.93**	31.42	**21.15**
Whisper-large-v3-turbo	61.93	**38.12**	35.06	**22.24**

Table 3: ASR performance on the Cypriot Greek test set before and after fine-tuning with normalized transcripts.

Model	WER (%)		CER (%)	
	Before	After	Before	After
Whisper-small	81.28	**51.39**	49.15	**29.85**
Whisper-medium	69.17	**43.64**	38.95	**28.22**
Whisper-large-v3	58.61	**35.70**	31.29	**22.30**
Whisper-large-v3-turbo	62.16	**36.73**	34.95	**21.88**

Table 4: ASR performance on the Cypriot Greek test set before and after fine-tuning using the original transcripts. Word Error Rate (WER) and Character Error Rate (CER) are reported.

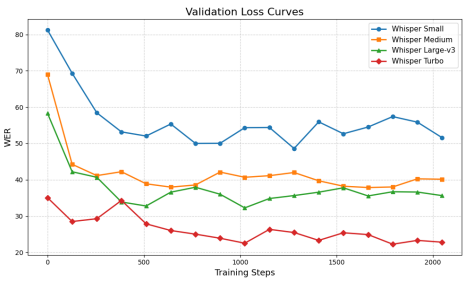


Figure 1: Validation WER curves across training steps for Whisper-small, Whisper-medium, Whisper-large-v3 and Whisper-large-v3-turbo.

As shown in Table 3 and Table 4, Whisper without fine-tuning achieves high WER scores for Cypriot Greek. We verify that this is due to the characteristics of the dialect as Whisper before fine-tuning achieves similar performance both on the original and the normalized transcripts, indicating that this is not due to transcription artifacts or rarely occurring characters and accents. Comparing Tables 3 and 4, normalization leads to consistent improvements for medium and large models (e.g., 5.8 absolute WER for Whisper-medium), while smaller models show less stable behavior, suggesting that normalization benefits models with sufficient capacity to exploit reduced orthographic variability.

[5] openai/whisper-small
[6] openai/whisper-medium
[7] openai/whisper-large-v3
[8] openai/whisper-large-v3-turbo
[9] lighteternal/wav2vec2-large-xlsr-53-greek
[10] https://github.com/huggingface/transformers/blob/main/examples/pytorch/speech-recognition/run_speech_recognition_seq2seq.py4

After fine-tuning with less than 20 hours of speech, all Whisper variants achieved significantly lower WER in both settings, with a $\sim 38\%$ and $\sim 40\%$ average relative WER reduction for the original and normalized transcripts, respectively. Among the variants, Whisper large v3 emerged as the superior model, achieving a WER of 33.93% and a CER of 21.15% on normalized transcripts. Performance consistently improves with model scale. However, although Whisper-large-v3-turbo is optimized for inference efficiency, it does not match the full large-v3 model in accuracy.

Figure 1 shows the validation WER across training steps. Larger models (*medium* and *large-v3*) not only achieve lower error rates but also converge more smoothly compared to *whisper-small*. While *whisper-small* exhibits notable fluctuations, peaking at 57% even late in training, *whisper-large-v3* stabilizes early, reaching its performance floor around step 1,000. This suggests that the higher parameter count in the large variant provides a more robust framework for dialectal transfer learning.

	Before		After	
Dataset	**WER**	**CER**	**WER**	**CER**
CY-original	84.07	48.75	**57.21**	**25.14**
CY-normalized	83.14	43.04	**46.96**	**17.88**

Table 5: ASR performance of wav2vec 2.0 XLS-R-greek on the Cypriot Greek test set before and after fine-tuning with original and normalized transcripts.

Table 5 reports the XLS-R-greek results. In the zero-shot scenario, WER exceeds 83% and CER is similarly high. Fine-tuning substantially reduces errors, with normalized transcripts leading to the best performance: 46.96% WER and 17.88% CER. Notably, XLS-R achieves WER reductions comparable to *whisper-medium* and competitive CER scores, but its WER remains higher than *whisper-large-v3*, suggesting that sequence-to-sequence modeling with an autoregressive decoder better captures word-level dependencies in dialectal speech.

Overall, the results highlight the importance of both model size and transcript normalization in ASR adaptation to low-resource dialects. Larger models capture dialectal variation more effectively, while normalization mitigates orthographically irrelevant variability. Fine-tuning on dialect-specific data is thus essential for achieving robust recognition of Cypriot Greek across different ASR architectures.

6. Previous studies

ASR models have achieved significant achievements in high-resource languages, English being the most represented instance of well-resourced language (Eisenstein et al., 2023; Pratap et al., 2024). Typical methods to deal with data scarcity issues involve transfer learning from high-resource languages, unsupervised pre-training with unpaired data, and synthetic data augmentation to improve model robustness and reduce overfitting (Conneau et al., 2020).

The most recent work on Greek dialects establishes a benchmark for six low-resource Greek varieties (Aivaliot, Aperathiot, Cappadocian, Cretan, Griko, and Messenian) and discusses the salient problems of limited data, orthographic differences, and language contact (Vakirtzian et al., 2024; Tsoukala et al., 2026). Initial experiments showed that state-of-the-art models and cross-lingual transfer methods struggle to adapt to these varieties. For instance, while ASR models handle SMG well (WER 11–14%), their performance drops drastically on Greek dialects, with WERs ranging from 28% (Cretan) to 100% (Griko). These difficulties arise from multiple factors, including significant phonological differences between SMG and the dialects, alongside variations in discourse register; dialectal speech is often recorded as oral speech rather than reading texts (Vakirtzian et al., 2024; Tsoukala et al., 2026). Consequently, varieties that exhibit higher phonological proximity to the high-resource SMG, such as the Cretan dialect, achieved the strongest performance.

7. Conclusions

This work presents the first ASR model for Cypriot Greek, addressing the challenge of data scarcity in ASR for low-resource dialects. Initial experiments indicate that Whisper struggles on non-standardized dialects (WER 60%) due to the variety-specific characteristics. This further highlights the challenge of recognizing non-standard dialects with out-of-the-box ASR models. Fine-tuning on less than 20 hours of dialect-specific speech substantially improves performance, reducing WER by 38% on original transcripts and 40% on normalized transcripts. The improvement demonstrates that even a relatively small amount of dialect-specific data can effectively adapt the model, with Whisper-large-v3 achieving the best results due to its greater capacity to capture complex speech patterns, a result that also benefits from Cypriot Greek's closer proximity to SMG as a southern Greek dialect. Moreover, these results highlight the importance of orthographic normalization when adapting ASR systems to dialectal varieties with inconsistent spelling conventions.

8. Future Work

In future iterations of this work, we plan to conduct a comprehensive comparative analysis between Cypriot Greek and Standard Modern Greek in order to quantify the "dialctal gap". Recent work on SMG (Paraskevopoulos et al., 2023) evaluated the performance of XLS-R on Standard Modern Greek (SMG) using the Mozilla Common Voice (CV) dataset. This dataset represents a "controlled" environment of read speech and consists of approximately 16.2 hours of total audio, with 12.27 hours specifically used for training. In this controlled read-speech environment, the model achieved a 29.33% WER.

In contrast, our fine-tuning results for Cypriot Greek reveal a notable performance disparity, which we hypothesize is due to the distinct phonological and morphological characteristics of the dialect, such as gemination and palatalization. Even after fine-tuning with dialect-specific data, our best-performing Whisper-large-v3 model achieved a WER of 33.93%, while our XLS-R-greek model reached 46.96% on normalized transcripts. By integrating these SMG benchmarks, we could: (i) identify the specific linguistic features of CG that contribute most to the increase in WER, (ii) evaluate whether specialized fine-tuning strategies, such as Mixed Multi-Domain Self-Supervision (M2DS2), can bridge this gap by anchoring target dialect representations to the source language, (iii) assess the cross-dialectal transferability of Whisper's zero-shot capabilities in comparison to fine-tuned self-supervised models like XLS-R.

9. Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.

10. Bibliographical References

Spyros Armostis. 2022. Η κυπριακή ελληνική τζ̌αι η γραφή της [=cypriot greek and its writing]. In Alexia Achilleos, Spyros Armostis, and Eleftheria Socratous, editors, ΑΠΟαποικιοΠΟΙΗΣΗ: Γλωσσοπλάσματα που μηχανές τζ̌αι πλάσματα *[=Decolonisation: Linguistic creations by machines and humans]*, pages 13–33. Ypogeia Skini, Limassol.

Spyros Armostis, Kyriaki Christodoulou, Marianna Katsoyannou, and Charalambos Themistocleous. 2014. Addressing writing system issues in dialectal lexicography: The case of cypriot greek. In Carrie Dyck, Tania Granadillo, Keren Rice, and Jorge Emilio Rosés Labrada, editors, *Dialogue on Dialect Standardization*, pages 23–38. Cambridge Scholars Publishing, Cambridge.

Thirunavukkarasu Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Jessica Chin, Elena Talevska, and Mark Antoniou. 2023. Speech-in-speech recognition is modulated by familiarity to dialect. pages 3113–3116.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of Interspeech*.

Nikolaos Contossopoulos. 1969. Συμβολή εις την μελέτην της Κυπριακής διαλέκτου [a contribution to the study of the cypriot dialect]. *Epetiris tu Kentru Epistimonikon Erevnon*, 3:87–109.

Jacob Eisenstein, Vinodkumar Prabhakaran, Christopher Rivera, Dorottya Demszky, and Divyansh Sharma. 2023. Md3: The multi-dialect dataset of dialogues. In *Proceedings of INTERSPEECH*, pages 4059–4063.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strotgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. In *North American Chapter of the Association for Computational Linguistics*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sushree Sangita Mohanty, Satya Ranjan Dash, and Shantipriya Parida, editors. 2024. *Applying AI-Based Tools and Technologies Towards Revitalization of Indigenous and Endangered Languages*. Studies in Computational Intelligence. Springer Singapore.

Brian Newton. 1972. *Cypriot Greek: Its Phonology and Inflections*. Mouton, The Hague.

Andreas Papapavlou. 2010. Language planning in action: searching for a viable bidialectal program. *Language Problems and Language Planning*, 34(2):120–140.

Georgios Paraskevopoulos, Theodoros Kouzelis, Georgios Rouvalis, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. 2023. Sample-efficient unsupervised domain adaptation of speech recognition systems: A case study for modern greek. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:286–299.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(1):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Statistical Service of Cyprus. 2011. *Population Census 2011: Population by Language*. Statistical Service of Cyprus, Nicosia.

Christos Symeonidis. 2006. Ιστορία της Κυπριακής διαλέκτου. *Istoria tis Kypriakis dialektou. [History of the Cypriot Dialect]*. Kentro Meleton Ieras Monis Kykkou, Nicosia.

Charalambos Themistocleous. 2010. Writing in a non-standard greek variety: Romanized cypriot greek in online chat. *Writing Systems Research*, 2(2):155–168.

Charalambos Themistocleous. 2019. Dialect classification from a single sonorant sound using deep neural networks. *Frontiers in Communication*, 4.

Stavroula Tsiplakou. 2014. How mixed is a 'mixed' system?: The case of the cypriot greek koiné. *Linguistic Variation*, 14(1):161–178.

Stavroula Tsiplakou, Andreas Papapavlou, Pavlos Pavlou, and Marianna Katsoyannou. 2006. Levelling, koineization and their implications for bidialectism. In Frans Hinskens, editor, *Language Variation - European Perspectives III: Selected Papers from the 3rd International Conference on Language Variation in Europe (Amsterdam, 23-25 June 2005)*, pages 265–276. John Benjamins, Amsterdam.

Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos, and Stella Markantonatou. 2026. Extending ASR evaluation resources for Modern Greek dialects. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 210–222, Rabat, Morocco. Association for Computational Linguistics.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech recognition for greek dialects: A challenging benchmark. In *Proceedings of Interspeech 2024*, pages 3974–3978.

Constantinos Yiangoullis. 2014. Θησαυρός της μεσαιωνικής και νεότερης κυπριακής διαλέκτου: Ερμηνευτικός, Ετυμολογικός, Φρασεολογικός και Ονοματολογικός *[Thesaurus of Medieval and Contemporary Cypriot Greek dialect: Interpretative, Etymological, Phraseological, and Onomastic]*. (n.p.), Λευκωσία.

Structural Divergence under Shared Language-Level Specification: Griko in Universal Dependencies

Stavros Bompolas[*1], Emanuela Pinna[*2], Josep Quer[2], Marika Lekakou[3], Stella Markantonatou[4]

[1]Archimedes/Athena RC, [2]Universitat Pompeu Fabra, [3]University of Ioannina,
[4]Institute of Language and Speech Processing/Athena RC
[1]Artemidos 1, 15125, Marousi, Greece, [2]Carrer de Roc Boronat 138, 08018, Barcelona, Spain,
[3]Department of Philology, 45110, Ioannina, Greece, [4]Aigialias 19 & Chalepa 3, 15125, Marousi, Greece
{s.bompolas, marks}@athenarc.gr, {josep.quer, emanuela.pinna}@upf.edu, mlekakou@uoi.gr

Abstract

Dialectal varieties pose major challenges for NLP resource development, especially when annotation frameworks are organized around standardized language specifications. In Universal Dependencies (UD), dialects without independent ISO codes are subsumed under the corresponding standard language and inherit its language-level documentation, validator settings, and grammatical inventories. This paper examines Griko, a Greek variety spoken in southern Italy that developed in relative isolation from the Modern Greek dialect continuum while remaining in long-term contact with local Italo-Romance varieties. We assess the consequences of this organizational structure through controlled parsing experiments comparing intra-dialectal training, cross-dialectal transfer from Standard Modern Greek (SMG), script-controlled transfer using romanized SMG, and contact-related cross-lingual transfer from Italian. Our results show that, before romanization, the Italian model even surpasses SMG on several UD metrics and that, although romanization substantially improves SMG-based transfer, performance still remains far below the intra-dialectal baseline. We argue that this persistent gap reflects the interaction between structural divergence and language-level validation constraints, a phenomenon we term *ISO-based validation coupling*. Through analyses of auxiliary systems, voice marking, and progressive constructions, we show how standard-centric validation architectures can constrain the representation of dialect-specific grammar. More broadly, the Griko case highlights the limitations of language-centric organization in UD and underscores the need for variety-sensitive mechanisms when extending universal annotation frameworks to structurally divergent dialects.

Keywords: Universal Dependencies, dialectal NLP, Griko, cross-dialectal transfer, morphosyntactic divergence

1. Introduction

Dialectal varieties pose fundamental challenges for NLP resource development, particularly when annotation frameworks are designed around standardized language varieties. Within the Universal Dependencies (UD) framework (Nivre et al., 2020; de Marneffe et al., 2021), dialects that lack independent ISO codes are typically subsumed under the standard language, implicitly assuming structural comparability and transferability. As a result, annotation decisions are constrained by shared inventories, validation rules, and morphosyntactic specifications defined at the language level.

We present a case study of Griko (ISO 639-3: `ell`; Glottolog: `apul1237`), a Modern Greek (MG) variety spoken in Salento (southern Italy). Griko has evolved in relative isolation from the MG dialect continuum and in prolonged contact with the local Italo-Romance varieties, resulting in extensive phonological and morphosyntactic divergence. Because Griko lacks its own ISO code, it is represented under the MG specification in UD, alongside Standard Modern Greek (SMG) and other MG varieties. As a consequence, dialect-specific grammatical categories cannot be independently specified or constrained without potentially affecting all Greek varieties grouped under the same language code, and vice versa. This organizational constraint has direct implications for annotation design, validation procedures, and computational modeling.

Griko has recently attracted attention in NLP as a severely under-resourced and endangered variety. Existing work has developed parallel corpora, POS-tagging resources, and speech processing datasets (Boito et al., 2018; Anastasopoulos et al., 2018; Vakirtzian et al., 2024), as well as alignment and correction methods for noisy or low-resource data (Rijhwani et al., 2020; Xie and Anastasopoulos, 2023). While these studies demonstrate the feasibility of applying NLP techniques to Griko (see also Ramponi, 2024), they primarily address task-specific modeling rather than foundational questions of syntactic resource design within a universal annotation framework.

In this paper, we address this gap by investigating the morphosyntactic and computational con-

*Equal contribution.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 315–326
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

sequences of incorporating Griko into UD under the MG specification. We argue that this standard-centric treatment creates both technical and theoretical challenges. To substantiate this claim, we evaluate four training configurations: (i) intra-dialectal training (Griko→Griko); (ii) cross-dialectal transfer from Standard Modern Greek (SMG→Griko); (iii) script-controlled cross-dialectal transfer using romanized SMG data (Romanized SMG→Griko); and (iv) contact-related cross-lingual transfer from Italian (Italian→Griko). The results show that Italian slightly outperforms non-romanized SMG on some UD metrics, suggesting that script alignment and contact-related overlap may offer limited advantages, although both transfer conditions remain weak overall. Crucially, even after orthographic differences are minimized, transfer from SMG to Griko remains extremely poor. This indicates that script discrepancies alone do not explain transfer failure; deeper morphosyntactic divergence, together with annotation-level constraints, plays a decisive role. Therefore, we examine specific annotation challenges, including auxiliary inventories, voice distinctions, and progressive constructions, to show how the shared UD specification obscures structural mismatches between Griko and SMG.

Our findings suggest that neither genealogical relatedness nor shared ISO coding is sufficient to guarantee structural compatibility for dependency parsing. Taken together, these results motivate a broader critique of UD as a language-centric framework. More broadly, we argue that standard-centric modeling practices and language-level validation architectures can obscure dialect-specific grammatical systems, thereby affecting both computational performance and faithful linguistic representation.

By combining empirical parsing experiments with annotation-level analysis, this study contributes to ongoing discussions on how universal frameworks, such as UD, can accommodate dialectal diversity without enforcing structural homogenization.

2. Background

This section situates the present study within the broader landscape of dialectal resource development in UD. We first outline how dialects are currently represented within the UD ecosystem and discuss emerging efforts to model non-standard varieties. We then review the status of MG in UD and examine how its organizational and validation structure affects the inclusion of additional varieties. This overview provides the necessary context for understanding the specific challenges posed by Griko.

2.1. Dialects in Universal Dependencies

The UD framework has made substantial progress in multilingual morphosyntactic annotation, currently covering more than 150 languages. Its cross-linguistically consistent design has enabled large-scale comparative research and transfer-based NLP modeling. However, extending UD to dialectal varieties exposes important theoretical and methodological challenges.

Recent work has increasingly focused on developing dialectal treebanks. According to a recent survey,[1] more than 30 dialectal UD treebanks have been identified. Treebanks have been created for dialects and non-standard varieties such as Egyptian Arabic (Maamouri et al., 2014), Norwegian dialects (Øvrelid et al., 2018; Kåsen et al., 2022), Occitan (Miletic et al., 2020), and Bavarian (Blaschke et al., 2024), among others. These initiatives reflect a growing recognition that dialectal resources are essential both for linguistic documentation and for building robust NLP systems capable of handling non-standard input.

Despite this progress, dialectal resources in UD remain constrained by the framework's language-based organization. UD's higher-level metadata are explicitly language-centered: the current language inventory states that information about language families and genera is mostly taken from WALS Online (Dryer and Haspelmath, 2013), a resource well suited to broad genealogical classification but too coarse to encode dialect-internal structural divergence. As a result, varieties without distinct ISO codes are often subsumed under a standard-language node and inherit language-level documentation and validation assumptions, including inventories of auxiliaries, morphological features, and consistency constraints. At the same time, UD also includes non-standard varieties that are treated as separate language entries with their own documentation, including cases that are neither separately represented in WALS nor assigned a distinct ISO code, such as Pomak (Markantonatou et al., 2023). The coexistence of these organizational strategies creates a structural tension between cross-dialectal comparability and dialect-specific adequacy, raising broader questions about how grammatical divergence can be represented within a shared language specification.

2.2. Modern Greek and its dialects in Universal Dependencies

The UD framework has been applied to SMG through two major treebanks: UD_Greek-GDT

[1] https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:wg1:worshop-submission-unidive.pdf

(Prokopidis and Papageorgiou, 2017), based on the Greek Dependency Treebank, and the more recent UD_Greek-GUD (Markantonatou et al., 2025), which adheres more closely to UD.v2 morphological guidelines. The two treebanks differ with respect to the registers they cover and the detail of modeling linguistic phenomena.

These resources serve as benchmarks for dialect-oriented NLP research. Newly developed treebanks for MG varieties are generally expected to align with SMG-based annotation guidelines to ensure compatibility within the UD ecosystem.

In recent years, five UD treebanks have been developed for MG dialects. Two focus on Cappadocian—UD_Cappadocian-AMGiC (Sampanis and Prokopidis, 2021) and UD_Cappadocian-TueCL (Vligouridou et al., 2024)—while three represent other dialects: UD_Greek-Cretan (Vakirtzian et al., 2025), UD_Greek-Lesbian (Bompolas et al., 2025), and UD_Greek-Messenian.

Cretan, Lesbian, and Messenian are annotated under the shared MG ISO code and follow the SMG-based morphosyntactic specification, maintaining substantial structural continuity with the standard variety.

Cappadocian, however, constitutes a different case. Like Griko, Cappadocian developed over centuries in relative isolation and under intense contact influence—primarily from Turkish (Dawkins, 1916). Crucially, Cappadocian is assigned its own ISO 639-3 code (`cpg`) and is therefore treated as a separate language in UD. Its treebanks are not bound by the MG validation constraints and can specify independent auxiliary inventories and feature systems without affecting other Greek treebanks.

Griko represents a structurally comparable case of long-term isolation and contact-induced change, yet it does not possess a separate ISO code. As a result, it must be incorporated under the MG language specification. This means that its annotation is subject to the same validation rules and feature inventories as SMG. Consequently, dialect-specific grammatical distinctions cannot be formalized without affecting all treebanks grouped under the same language code.

Incorporating Griko into UD is, therefore, not simply a matter of adding a new dataset. It requires reconciling dialectal morphosyntactic divergence with a validation architecture designed for the standard language.

2.3. The Griko Variety

Griko is a Greek variety spoken in Salento, in the south-eastern extremity of Italy, and has been attested at least since the Middle Ages. Its precise origin has long been debated. The main hypotheses summarized in Manolessou (2005, and references therein) converge into three principal theories: (a) Griko reflects an unbroken Greek presence in Southern Italy dating back to the Magna Graecia colonies of the 8th–7th centuries BCE; (b) it originated during the Byzantine presence in Southern Italy between the 6th and 11th centuries CE; or (c) like most MG dialects (with the exception of Tsakonian), it participated in the broader evolution from Hellenistic Koiné to Late Medieval Greek through continuous contact between Southern Italy and the Greek-speaking world, with archaic traits reflecting preserved Koiné features. Today, theories (b) and (c) are generally regarded as the most plausible.

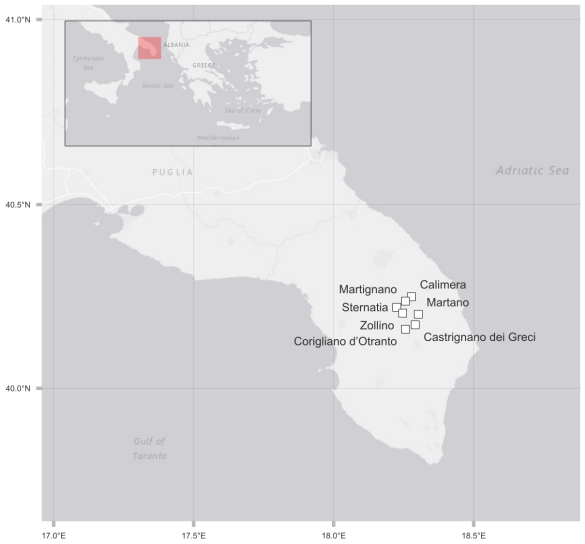


Figure 1: Geographic distribution of the remaining Griko-speaking communities, concentrated in seven towns of Salento (Apulia, Southern Italy).

Regardless of its precise historical origin, Griko has developed for centuries in sustained contact with Salentino, the local Italo-Romance variety, and later with Standard Italian. This prolonged bilingual setting has resulted in extensive contact-induced change across phonology, lexicon, and morphosyntax. Several core grammatical domains—including auxiliary formation, voice marking, and progressive constructions—show restructuring aligned with Italo-Romance patterns. These developments are particularly relevant for syntactic annotation, as they diverge systematically from SMG.

From a demographic perspective, Griko is classified by UNESCO as a severely endangered language (Moseley, 2012). Intergenerational transmission has largely ceased since the Second World War, and the language is today mainly spoken by elderly speakers. Its remaining speech communities are concentrated in seven towns in Salento (Figure 1)—Calimera, Castrignano dei Greci, Corigliano d'Otranto, Martano, Martignano, Sternatia, and Zollino—where vitality varies. Sociolinguistically, Griko has historically existed in a diglossic relation-

ship with Salentino and Italian, functioning as the lower-prestige code largely restricted to informal and domestic domains (see Pellegrino, 2021 for a comprehensive diachronic and synchronic account).

Griko has traditionally been transmitted orally and still lacks an officially standardized orthography. Written documentation, primarily in the form of poetry and narrative texts, began to appear in the late 19th century and consistently employs a Latin-based alphabet. The absence of orthographic standardization introduces additional challenges for corpus development and NLP processing, especially when resources are integrated into infrastructures primarily designed for Greek-script varieties.

Taken together, Griko represents a case of long-term contact-driven divergence within the MG dialect continuum. Its structural development under conditions of isolation, bilingualism, and sociolinguistic marginalization renders it particularly informative for examining how standard-centric annotation frameworks accommodate—or fail to accommodate—deeply divergent dialectal systems.

3. Experimental Design

The issues discussed above motivate an empirical evaluation of transfer-based dependency parsing within UD. To this end, we conduct controlled parsing experiments comparing intra-dialectal, cross-dialectal, script-controlled, and contact-related cross-lingual training conditions. This design allows us to test whether shared ISO-based grouping actually yields cross-variety parsing compatibility, or whether morphosyntactic divergence between Griko and SMG constrains transfer despite genealogical relatedness.

3.1. Data

Griko Dataset. The Griko data used in this study originate from the Palumbo corpus (Tommasi, 1998), a collection of 114 narrative texts originally recorded and documented by Vito Domenico Palumbo (1854–1928). The corpus was recently digitized and enriched with Italian translations and partial gold part-of-speech annotation (Anastasopoulos et al., 2018).

For the present study, a subset of 639 sentences from the Palumbo corpus was manually annotated and converted into CoNLL-U format by a trained linguist with specialized expertise in Griko grammar. Annotation decisions followed the UD.v2 guidelines and were aligned, where structurally possible, with the conventions adopted in the UD_Greek-GUD treebank. However, full alignment with SMG was not always feasible due to systematic morphosyntactic divergence. In such cases, annotation choices prioritized preserving grammatical distinctions internal to Griko.

The Griko treebank should currently be considered under development. For the experiments reported here, validation was restricted to format-level compliance using the official UD CoNLL-U validator.[2] Language-specific morphological and syntactic validation rules were not fully enforced, as these are presently defined only for SMG under the shared ISO code.

The dataset was partitioned into training and test splits; for the trainable conditions, Stanza automatically created the development split from the training portion.

Standard Modern Greek Dataset. For SMG, we use the UD_Greek-GUD treebank (Markantonatou et al., 2025). UD_Greek-GUD (GUD) is the most recent UD treebank for SMG. The corpus comprises fiction texts as well as material from online sources that reflect colloquial SMG.

This dataset represents the standard-based morphosyntactic specification to which additional MG varieties must formally align within the UD framework. It therefore serves as the baseline source for the cross-dialectal transfer experiments reported in this study. However, SMG is used here primarily as a convenient reference variety and comparison baseline, rather than as the most historically appropriate Greek point of comparison for Griko.

For the purposes of this study, we used the most current version of the GUD training set. In trainable conditions, Stanza automatically derived the development split from the training data.

Romanized SMG Dataset (Script-Control Condition). To isolate the effect of script differences between SMG (Greek alphabet) and Griko (Latin alphabet), we created a romanized version of the GUD treebank.

Romanization was performed automatically using a rule-based transliteration pipeline (uroman; Hermjakob et al., 2018)[3] applied directly to the CoNLL-U files. The procedure preserves tokenization, morphological features, and dependency structure while transforming only the orthographic representation into Latin script.

This dataset allows us to evaluate whether transfer degradation from SMG to Griko is primarily due to script mismatch or instead reflects deeper structural divergence.

[2] https://github.com/UniversalDependencies/tools/blob/master/udtools/README.md#the-official-udconll-u-validator

[3] https://github.com/isi-nlp/uroman

Dataset / Model	Split	# Sentences	# Tokens
Griko	train	575	6,992
Griko	test	64	783
GUD (SMG)	train	1,071	15,931

Table 1: Dataset statistics for the experimental setup. The Romanized GUD dataset corresponds to a script-transformed version of UD_Greek-GUD and therefore shares identical sentence and token counts. The Italian condition is based on Stanza's official pretrained combined Italian model; see the main text for further details.

Italian Model. To assess whether a contact-related language provides an advantage, we additionally evaluate cross-lingual transfer from Italian using Stanza's official pretrained Italian dependency parsing pipeline.

According to the Stanza documentation,[4] the default Italian package is a combined model trained on multiple UD Italian treebanks, namely ISDT, VIT, PoSTWITA, and TWITTIRO. Because this condition relies on a released pretrained package rather than on a model retrained under the same train/dev split regime as our other trainable conditions, it should be interpreted as a comparison baseline rather than as a fully controlled parallel training condition.

This condition is included as a convenient Italo-Romance comparison point rather than as a direct proxy for the local Italo-Romance varieties historically in contact with Griko.

3.2. Methods

Training Configurations. We evaluate four training configurations:

1. **Intra-dialectal transfer (Griko→Griko)**: The model is trained and evaluated exclusively on Griko data. This condition establishes an upper-bound baseline for parsing performance given available training data.
2. **Cross-dialectal transfer (SMG→Griko)**: The model is trained on SMG (UD_Greek-GUD) and evaluated on Griko. This condition tests the hypothesis that structural similarity within the MG ISO grouping enables knowledge transfer.
3. **Script-controlled transfer (Romanized SMG→Griko)**: The model is trained on romanized SMG and evaluated on Griko. This configuration controls for orthographic mismatch while preserving the SMG morphosyntactic specification.
4. **Cross-lingual (contact-related) transfer (Italian→Griko)**: The model is the official pretrained Stanza combined Italian model and is evaluated directly on the Griko test set. This condition tests whether a Romance baseline linked to the contact ecology of Griko offers a more informative transfer source than SMG.

All models are evaluated on the same held-out Griko test set.

[4] https://stanfordnlp.github.io/stanza/combined_models.html

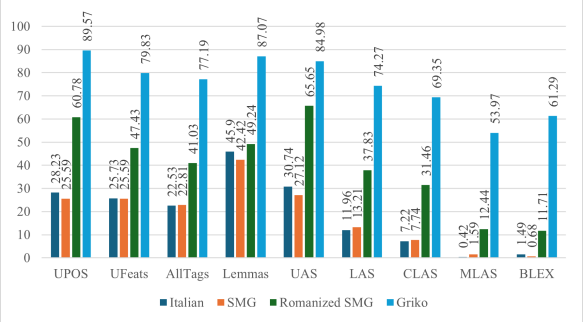


Figure 2: Dependency parsing performance (F1) on the Griko test set under four experimental conditions: intra-dialectal training, cross-dialectal transfer from SMG, script-controlled transfer using romanized SMG, and contact-related cross-lingual transfer from Italian.

Model Architecture. All experiments were conducted using the Stanza neural dependency parser (Qi et al., 2020). For SMG training, we use Greek-specific contextual embeddings (nlpaueb/bert-base-greek-uncased-v1) (Koutsikakis et al., 2020), which are appropriate for Greek-script data. For Griko and romanized SMG training, we use multilingual contextual embeddings (jhu-clsp/mmBERT-base) (Marone et al., 2025), which are better suited to Latin-script input and cross-dialectal generalization. For the Italian baseline, we use the official pretrained Stanza combined Italian model distributed with Stanza. Hyperparameters followed Stanza default configurations.

Evaluation. Parsing performance is evaluated using the official UD parsing scorer.[5] We report standard UD metrics, expressed as F1 scores.

4. Results

Figure 2 reports dependency parsing performance (F1 scores) across the four training configurations.

Cross-dialectal Transfer (SMG→Griko). When a model trained on SMG is evaluated on Griko, performance drops sharply across all metrics. Labeled Attachment Score (LAS) reaches only 13.21, while

[5] https://github.com/UniversalDependencies/tools/blob/master/udtools/README.md#the-official-ud-parsing-scorer

MLAS and BLEX—metrics sensitive to the interaction between morphology and syntax—approach zero (1.59 and 0.68, respectively).

Two factors may contribute to this degradation. First, there is substantial script distance between the varieties: SMG is written in the Greek alphabet, whereas Griko uses a Latin-based orthography. Because modern parsers rely on contextualized subword representations, orthographic mismatch introduces immediate divergence at the input level. Second, SMG and Griko differ in key morphosyntactic domains, including auxiliary inventories, voice marking, and progressive constructions. The near-total collapse of MLAS indicates that morphological feature prediction fails almost entirely under direct transfer, suggesting incompatibility at the level of feature systems and syntactic structuring rather than simple lexical mismatch.

Compared with Italian, SMG remains slightly better on LAS, CLAS, and MLAS, but weaker on several metrics before romanization. This contrast is important for our assumptions: relatedness to Greek does not automatically overcome script-induced and structural mismatch, while contact-related Romance similarity can help locally without yielding stronger syntactic transfer.

Script-Controlled Transfer (Romanized SMG→Griko). To disentangle orthographic effects from structural divergence, we train on a romanized version of SMG while preserving its morphosyntactic annotation.

Performance improves substantially relative to direct transfer: LAS increases from 13.21 to 37.83, UAS from 27.12 to 65.65, and UPOS to 60.78. These gains confirm that script discrepancy is a significant source of transfer degradation, as orthographic alignment facilitates lexical and subword generalization across varieties.

Romanized SMG also clearly outperforms both original-script SMG and Italian on every reported metric, showing that script normalization matters much more than the choice between those two unnormalized donor conditions. However, performance remains well below the intra-dialectal baseline. LAS reaches 37.83 compared to 74.27 in the Griko→Griko condition, and MLAS remains low at 12.44. The remaining gap indicates that structural divergence—not script difference alone—constitutes the primary limitation in cross-dialectal transfer.

Contact-Related Cross-Lingual Transfer (Italian→Griko). When the official pretrained Italian Stanza combined model is evaluated on Griko, performance is likewise poor overall: LAS reaches 11.96, CLAS 7.22, MLAS 0.42, and BLEX 1.49. At the same time, Italian slightly outperforms non-romanized SMG on UPOS, UFeats, Lemmas, UAS, and BLEX. This suggests that orthographic compatibility and contact-related lexical overlap can help some surface-oriented predictions; however, Italian does not yield better syntax-sensitive transfer than Greek.

Intra-Dialectal Training (Griko→Griko). When trained and evaluated exclusively on Griko data, the model achieves robust performance despite the relatively limited dataset size. This result shows that the Griko dataset forms a learnable and internally consistent system. The low performance observed under cross-dialectal transfer cannot therefore be attributed solely to data sparsity or annotation instability, but reflects genuine cross-dialectal divergence.

Summary. The four configurations reveal a consistent pattern:

1. Direct transfer from SMG to Griko yields near-complete failure.
2. Transfer from Italian is also poor overall, although it slightly outperforms non-romanized SMG on some metrics.
3. Script normalization substantially improves performance and clearly surpasses both original-script SMG and Italian, but leaves a large residual gap.
4. Intra-dialectal training produces strong results.

Taken together, these findings indicate that neither shared Greek ISO grouping nor a contact-related Romance baseline guarantees structural interoperability for dependency parsing. The fact that Italian can outperform non-romanized SMG on some metrics prior to romanization is consistent with the role of orthography and contact-related overlap. Crucially, the gap between Romanized SMG and Griko-to-Griko training shows that deeper morphosyntactic divergence remains the principal obstacle to transfer.

5. Annotation Constraints and Dialect-Specific Divergence

The experimental results indicate that cross-dialectal transfer between SMG and Griko remains highly limited, even after controlling for script differences. While romanization substantially improves performance, a large gap persists between cross-dialectal and intra-dialectal training. This residual gap points to deeper structural divergence rather than surface-level orthographic mismatch. In this section, we examine how this divergence interacts with UD's organizational principles, particularly its language-level validation framework.

5.1. Griko under ISO-Based Validation Coupling

Within UD, Griko is not assigned an independent ISO 639-3 code and is therefore classified under MG. As a result, it cannot be represented as a separate language-level treebank, but must conform to the MG specification, including its inventories of auxiliaries, morphological features, and language-specific validation checks enforced by the official UD validator.

We refer to this configuration as *ISO-based validation coupling*: varieties grouped under the same ISO code are validated against a shared grammatical specification, irrespective of internal structural divergence. Validation thus operates at the level of the language code, not at the level of individual varieties.

This arrangement has concrete consequences for annotation.

First, auxiliary inventories are defined and validated at the language level. Auxiliaries that exist in Griko but not in SMG cannot be treated as variety-specific elements without modifying the shared MG specification. Conversely, auxiliaries shared across varieties but realized in different scripts introduce additional complexity. In the Griko case, the use of the Latin alphabet prevents direct orthographic ambiguity; however, for dialects written in the Greek alphabet, lexical verbs could potentially be validated as auxiliaries without triggering language-level inconsistencies.

Second, morphological feature inventories are likewise shared across all MG treebanks. Griko maintains a systematic distinction among `Voice=Act`, `Voice=Mid`, and `Voice=Pass`, whereas SMG does not preserve a comparable tripartite opposition. Because feature values are validated at the language level, the distribution of Voice=Mid cannot be restricted to Griko alone. Dialect-specific grammatical contrasts therefore cannot be formally encoded without simultaneously altering the specification for all MG varieties.

In what follows, we illustrate these dynamics through three case studies—auxiliary formation, voice marking, and progressive constructions—which demonstrate how Griko reorganizes core morphosyntactic domains in ways that challenge language-level validation assumptions within UD.

5.2. Case Study I: Dialect-Specific Auxiliaries

Griko has developed auxiliaries that do not exist in SMG, notably *èrkome* and *ènna*.

***èrkome* as Eventive Passive Auxiliary** *Èrkome* functions as an eventive passive auxiliary in both present and past contexts, reflecting contact-induced grammaticalization from Italo-Romance.

(1) èrkete fameno
come.IND.PRS.3SG.MID eat.PTCP.PASS.NOM.M.SG
'He is (being) eaten.' (Baldissera, 2013, 43)

***ènna* as Modal Auxiliary** *Ènna* is a deontic modal auxiliary derived from a grammaticalized construction corresponding to SMG *ekhi na*. Unlike SMG, Griko has reanalyzed this construction into an invariable auxiliary. Because it encodes modality beyond TAM distinctions, it is annotated with `VerbType=Mod`.

(2) fseri ti
know.IND.PRS.2SG.ACT what

ènna kamin
AUX do.IND.PRS.PFV.2SG.ACT
'Do you know what you should do?'
(Tommasi, 1998, Tale 27)

These auxiliaries exemplify structural developments absent from SMG. Under ISO-based validation coupling, however, incorporating them requires modification of the shared MG auxiliary inventory, affecting all treebanks under the same code.

5.3. Case Study II: Middle vs. Passive Voice

Griko exhibits a systematic distinction between middle and passive constructions. Unlike SMG, which employs a unified non-active morphology, Griko restricts inherited non-active morphology to middle-type uses (reflexive, anticausative, unaccusative). Passive meaning is instead expressed periphrastically via *ime*/*èrkome* + perfect passive participle.

(3a) plènome
wash.IND.PRS.1SG.MID
'I wash myself.'

(3b) èrkome plimeno
come.IND.PRS.1SG.MID wash.PTCP.PERF.PASS.NOM.M.SG
'I am being/getting washed.'

In these periphrastic constructions, no dedicated `Voice=Pass` morphology appears on the auxiliaries themselves; passive interpretation is signaled by the participle. Although passive voice could in principle be inferred from interpretation, we encode it explicitly at the syntactic level by using the dependency relation `aux:pass`.

A further complication arises from the fact that, in older Griko and in some contemporary varieties, BE-perfect (4a) and BE-passive constructions (4b) are morphologically identical:

(4a) ime ertomeno / ertomeni
be.IND.PRS.1SG come.PTCP.PERF.PASS.NOM.M.SG / F.SG
'I have arrived (m/f).'

(4b) ime sfammeno / sfammeni
be.IND.PRS.1SG kill.PTCP.PERF.PASS.NOM.M.SG / F.SG
'I am killed (m/f).'

Following the approach adopted in UD_Greek-GUD, participles in *-menos* are annotated as passive verbal forms (Markantonatou et al., 2025). Since no morphological distinction differentiates perfect from passive in these constructions, the contrast is expressed syntactically: the dependency relation `aux:pass` is used in (4b) to mark passive interpretation explicitly. Annotating the participle in (4a) as a middle formation would remove the need for this syntactic distinction; however, such an analysis would depart from the generalizations established in GUD and compromise cross-treebank consistency.

In contemporary varieties of Griko, perfect participles tend to be invariable in *-mena*, whereas passive participles retain agreement.

5.4. Case Study III: Progressive and Pseudo-Coordination Constructions

Griko exhibits pseudo-coordination structures of the form V1 + coordinator + V2, conveying aspectual or discourse-related functions. Following SMG annotation practice, such constructions are annotated with `conj` (or `parataxis` when the coordinator is absent), and semantically bleached V1 verbs receive `VerbType=Mod`.

A particularly revealing case involves the progressive construction with *ste(o)* 'to stay', which appears in three stages of grammaticalization:

(5a) Inflected *steo* + ce + V2
ìstike ce piske
stay.IND.PST.IPFV.3SG CONJ fish.IND.PST.IPFV.3SG
'He was fishing.' (Tommasi, 1998, Tale 107)

(5b) Invariable *ste* + ce + V2
ste ce piske
stay.DEFAULT CONJ fish.IND.PST.IPFV.3SG
'He was fishing.'

(5c) Invariable *ste* + V2
ste piske
stay.DEFAULT fish.IND.PST.IPFV.3SG
'He was fishing.'

In cases (5a) and (5b), we maintain coordination analysis. In (5c), the absence of inflection and conjunction signals full grammaticalization, and *ste* is annotated as `AUX`, in accordance with UD guidelines.

A parallel pseudo-coordination pattern occurs with *pianno* 'to grab':

(6a) pianni ce lei
grab.IND.PRS.3SG CONJ say.IND.PRS.3SG
'He grabs and says.'

(6b) pianni èftase èssu-ti
grab.IND.PRS.3SG arrive.IND.PST.PFV.3SG inside-she.GEN.F.SG
'He grabbed and arrived at her place.'
(Tommasi, 1998, Tale 46)

Unlike *ste*, however, *pianni* does not encode aspectual information and is therefore analyzed as a lexical verb in coordination or parataxis.

5.5. Summary

Across auxiliary formation, voice marking, and progressive constructions, Griko exhibits systematic morphosyntactic reorganization shaped by prolonged contact and isolation. These are not superficial lexical differences but structural divergences affecting core grammatical domains.

The severe degradation observed in cross-dialectal transfer thus reflects genuine grammatical incompatibility. ISO-based validation coupling ensures administrative coherence within UD, but it does not guarantee structural compatibility across dialects. Incorporating Griko into UD therefore requires more than data integration; it demands reconsideration of how dialect-level grammatical divergence is accommodated within language-level validation architectures.

6. Discussion: Toward More Flexible Treatment of Dialects in UD

The Griko case points to a broader issue for UD: how can the framework adequately represent structurally coherent lects that are dialectal, may lack a standardized parent language, or do not have an ISO code of their own? Our results suggest that this is not a marginal technical issue, but a central methodological challenge for the representation of dialectal diversity in universal annotation frameworks.

UD does not need to abandon its universal layer to accommodate dialects more adequately, but it does require greater flexibility in how varieties are documented, validated, and organized. One practical step would be to introduce explicit *variety profiles* under a shared language node, coupled with layered validation in which a treebank is checked first against universal and language-level UD constraints and then against variety-sensitive rules at the treebank or dialect level. This would preserve comparability across related treebanks without imposing the same decisions on all treebanks grouped under a single language code. More explicit variety-level documentation would also help separate genealogical affiliation from annotation governance, especially since UD's higher-level classification draws heavily on WALS, whose language-centered family and genus groupings are too coarse to capture dialect-internal structural divergence.

The problem is even sharper for varieties that lack both a standardized parent language and an independent ISO code. In such cases, rep-

resentational adequacy should not depend exclusively on administrative languagehood. From this perspective, a more flexible organizational basis—potentially drawing on resources such as Glottolog (Hammarström et al., 2026), which is designed to catalogue documented *languoids* below the standardized-language level—could offer a more flexible and comprehensive classificatory basis than ISO- or WALS-based organization alone.

At the same time, UD already shows that some flexibility is possible: Cappadocian and Pomak are treated as separate languages rather than being folded into broader genealogical groupings, even where ISO coding is absent, as with Pomak. This suggests that the issue is not the lack of mechanisms, but the absence of a consistent principle for when a dialect receives its own profile versus remaining under a broader language-level specification.

More generally, the Griko case highlights a tension within UD between cross-dialectal comparability and dialect-specific adequacy. Our results show that this tension is not only descriptive but also computationally consequential: neither shared ISO grouping nor genealogical relatedness guarantees effective transfer, as illustrated by the contact-related Italo-Romance baseline slightly outperforming non-romanized SMG. A more flexible approach would allow dialectal lects to retain a shared universal annotation layer while receiving their own formal profiles when justified.

7. Conclusions and Future Work

This study examined the incorporation of Griko into the UD framework under the MG specification. Through controlled parsing experiments, we showed that transfer to Griko remains severely limited under both cross-dialectal and cross-lingual conditions. Direct transfer from SMG performs extremely poorly, even relative to an Italian baseline. Although script normalization substantially improves performance in the SMG condition, a large residual gap persists relative to intra-dialectal training. These results indicate that orthographic mismatch explains only part of the degradation, whereas deeper morphosyntactic divergence remains the principal obstacle to transfer. We further substantiate this interpretation through annotation-level analysis.

We argued that this divergence interacts with what we term *ISO-based validation coupling*: varieties grouped under the same ISO code are validated against a shared grammatical specification even when their internal structures differ substantially. In Griko, auxiliary inventories, voice distinctions, and progressive constructions reveal systematic reorganization in core grammatical domains. Because validation operates primarily at the language level, dialect-specific categories cannot be independently constrained without potentially affecting all MG treebanks. More broadly, the Griko case highlights the limits of UD's language-centric organization. Its higher-level classification, partly based on WALS, is useful for broad genealogical organization but too coarse to capture dialect-internal structural divergence.

More generally, these findings suggest that standard-centric modeling practices may overestimate cross-dialectal transferability and obscure dialect-specific grammatical systems. Griko is not computationally well approximated by genealogically related SMG; indeed, before romanization, the contact-related Romance baseline slightly surpasses SMG on some UD metrics. The strongest performance is achieved when the model is trained on Griko itself, indicating that the variety constitutes a coherent grammatical system in its own right. Taken together, these results show that genealogical relatedness and shared grouping do not guarantee syntactic compatibility for dependency parsing (Lin et al., 2019; Faisal and Anastasopoulos, 2022), particularly in contact settings and where script differences are involved (see Vakirtzian et al., 2024 and Tsoukala et al., 2026 for similar patterns in Greek ASR). Empirical evaluation, combined with annotation-level analysis, is therefore essential when extending universal annotation frameworks to structurally divergent dialects.

Incorporating dialects into UD thus requires more than simply adding new treebanks under an existing language label. It requires mechanisms that preserve the universal annotation layer while allowing dialect-specific structure to be documented and validated on its own terms. As discussed above, possible directions include explicit variety profiles, layered validation, clearer separation between genealogical affiliation and annotation governance, and more flexible organizational bases for lects without independent ISO codes.

Future work will expand the Griko treebank, compare Griko more systematically with other (MG) dialect treebanks, and test whether the observed pattern holds across additional parsing architectures and transfer settings. Further work within UD is also needed to explore how variety-sensitive validation mechanisms can be implemented in practice without sacrificing interoperability.

Ultimately, the Griko case shows that incorporating dialects into UD requires more than adding new data: it requires rethinking how a language-level validation architecture can accommodate structural diversity within genealogically related varieties, and how standard-centric design decisions shape the representation and computational modeling of dialect-specific grammatical systems.

8. Limitations

Several limitations should be acknowledged.

First, the Griko treebank is still under development and relatively small in size. Although intra-dialectal transfer yields robust performance, limited data may affect the stability of fine-grained morphosyntactic evaluation and the generalizability of absolute scores. Expanding the corpus will allow for more precise assessment of structural phenomena and model behavior.

Second, validation of the Griko data was restricted to format-level compliance using the official UD CoNLL-U validator. Language-specific morphological and syntactic validation rules were not fully enforced, as these are currently defined only for SMG under the shared ISO code. While this reflects the structural issue examined in the paper, it also means that some annotation-level inconsistencies may remain undetected.

Third, the experimental setup relies on a single parsing architecture (Stanza) and specific contextual embedding choices. Although these configurations are standard and appropriate for the respective scripts, results may vary under alternative modeling approaches. Future work should evaluate whether the observed transfer limitations persist across architectures and multilingual pretraining strategies.

Fourth, the Italian condition uses the official pretrained Stanza combined model rather than a model trained by us on a manually selected Italian treebank split. This makes the baseline informative as a practical contact-related comparison, but not fully controlled in the same way as the SMG, Romanized SMG, and Griko conditions.

Finally, this study focuses on a single non-standard variety shaped by long-term isolation and language contact. While the analysis highlights structural challenges arising from ISO-based validation coupling, further comparative work across additional dialects will be necessary to determine the broader generalizability of the findings.

These limitations do not undermine the central empirical observation—that structural divergence substantially constrains cross-dialectal transfer—but they delimit the scope within which the present conclusions should be interpreted.

9. Data Availability

All data, models, and scripts used in this study are publicly available at https://osf.io/acf2x/overview. The repository includes the Griko UD-style treebank, the original UD_Greek-GUD dataset and its romanized version used in the script-controlled experiments, the corresponding Python romanization script, and the trained Stanza dependency parsing models for all experimental configurations. For the Italian baseline, we report the evaluation scores obtained with the official pretrained Stanza combined Italian model.

10. Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

We are especially grateful to the Griko speakers for their time, generosity, and invaluable help throughout this work.

11. Bibliographical References

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-Speech Tagging on an Endangered Language: a Parallel Griko-Italian Resource. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2529–2539, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Valeria Baldissera. 2013. *Il dialetto grico del Salento: elementi balcanici e contatto linguistico*. Ph.D. dissertation, Università Ca' Foscari Venezia.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, and Laurent Besacier. 2018. A small Griko-Italian speech translation corpus. *CoRR*, abs/1807.10740. ArXiv: 1807.10740.

Stavros Bompolas, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2025. Crossing dialectal boundaries: Building a treebank for the dialect of lesbos through knowledge transfer from standard Modern Greek. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 39–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Richard MacGillivray Dawkins. 1916. *Modern Greek in Asia Minor: a study of the dialects of Sílli,*

Cappadocia and Phárasa with grammar, texts, translations and glossary. Cambridge University Press, Cambridge.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-Inspired Adaptation of Multilingual Models to New Languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. glottolog/glottolog: Glottolog database 5.3.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, pages 110–117, New York, NY, USA. Association for Computing Machinery. Event-place: Athens, Greece.

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022. The Norwegian Dialect Corpus Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4827–4832, Marseille, France. European Language Resources Association.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2348–2354, Reykjavik, Iceland. European Language Resources Association (ELRA).

Io Manolessou. 2005. The Greek Dialects of Southern Italy: An Overview. *Cambridge Papers in Modern Greek*, 13:103–125.

Stella Markantonatou, Vivian Stamou, Stavros Bompolas, Katerina Anastasopoulou, Irianna Linardaki Vasileiadi, Konstantinos Diamantopoulos, Yannis Kazos, and Antonios Anastasopoulos. 2025. VMWE identification with models trained on GUD (a UDv.2 treebank of Standard Modern Greek). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 14–20, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Stella Markantonatou, Nicolaos Th. Constantinides, Vivian Stamou, Vasileios Arampatzakis, Panagiotis G. Krimpas, and George Pavlidis. 2023. Methodological issues regarding the semi-automatic UD treebank creation of under-resourced languages: the case of Pomak. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 27–35, Washington, D.C. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. _eprint: 2509.06888.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Christopher Moseley. 2012. *The UNESCO atlas of the world's languages in danger: context and process*. World Oral Literature Project. OCLC: 811994673.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis

Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Manuela Pellegrino. 2021. *Greek Language, Italian Landscape: Griko and the Re-storying of a Linguistic Minority*. Harvard University Press / Center for Hellenic Studies.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alan Ramponi. 2024. Language Varieties of Italy: Technology Challenges and Opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Konstantinos Sampanis and Prokopis Prokopidis. 2021. Asia Minor Greek in Contact (AMGiC): Towards a dialectal Treebank comprising contact-induced grammatical changes. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 86–95, Sofia, Bulgaria. Association for Computational Linguistics.

Salvatore Tommasi. 1998. *Io' mia forà… Fiabe e Racconti della Grecìa Salentina. Dai quaderni (1883–1912) di Vito Domenico Palumbo*. Ghetonia, Galatina, Lecce.

Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos, and Stella Markantonatou. 2026. Extending ASR Evaluation Resources for Modern Greek Dialects. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 210–222, Rabat, Morocco. Association for Computational Linguistics.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech Recognition for Greek Dialects: A Challenging Benchmark. In *Interspeech 2024*, pages 3974–3978. ISCA.

Eleni Vligouridou, Inessa Iliadou, and Çağrı Çöltekin. 2024. A Treebank of Asia Minor Greek. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1715–1721, Torino, Italia. ELRA and ICCL.

Ruoyu Xie and Antonios Anastasopoulos. 2023. Noisy Parallel Data Alignment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1501–1513, Dubrovnik, Croatia. Association for Computational Linguistics.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA Treebank of Spoken Norwegian Dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Digital Preservation of Aromanian Through Knowledge Management and Automatic Speech Recognition Evaluation

Marija Pendevska, Hristina Nastevska
Ss. Cyril and Methodius University of Skopje
Boulevard Goce Delchev 9, Skopje 1000, North Macedonia
{m.pendevska, hristinanastevska}@gmail.com

Abstract

This paper presents a knowledge management framework for the digital preservation of Aromanian, an endangered Eastern Romance language spoken across the Balkans, combined with the first systematic evaluation of automatic speech recognition (ASR) models on Aromanian dialectal speech. The proposed three-module framework encompasses localization of existing resources, distribution through digital platforms, and creation of new linguistic content through technological innovation. Empirical analysis of knowledge management practices among 176 respondents validates the framework design, revealing that information quality (28.3%) and personal utility (24.1%) are the most valued criteria for knowledge sharing, while digital information seeking (27.1%) is the dominant behaviour. Within this framework, we evaluate models from the OpenAI Whisper family across three sizes (medium, large-v2, large-v3) and multiple language settings on two Aromanian varieties: Gramosteanj and Crushova. All configurations yield word error rates (WER) above 88%, with character error rates (CER) as low as 34% under optimal conditions, indicating partial phonotactic capture despite word-level failure. The Latin language setting with large-v3 consistently achieves the best results. Romanization of non-Latin script output substantially reduces CER, confirming script mismatch as a major error source. These findings underscore the limitations of current pretrained ASR models for endangered languages and the need for dedicated resources and adaptation strategies within a broader language preservation framework.

Keywords: Aromanian, endangered language preservation, automatic speech recognition, knowledge management, Whisper, dialectal variation, digital preservation

1. Introduction

The Aromanian language represents a critical case in the preservation of endangered languages. With only a few thousand predominantly adult speakers remaining, Aromanian faces imminent extinction despite official recognition as a minority language in North Macedonia (Kahl, 2008). This precarious situation reflects broader global challenges: approximately 40% of the world's 6,000–7,000 languages are considered endangered (Moseley, 2010), and speakers of regional and underrepresented varieties remain largely excluded from modern speech and language technologies.

The need for systematic digital preservation of Aromanian extends beyond traditional documentation approaches that focus primarily on linguistic analysis. As Crystal (2000) emphasizes, language revitalization requires not only documentation but also the capacity for active use in contemporary contexts. This study addresses this gap by proposing a comprehensive framework that treats language preservation as a knowledge management challenge, encompassing collection, storage, distribution, accessibility, and continuous creation of knowledge (Pendevska, 2019).

Recent advances in automatic speech recognition (ASR), driven by large-scale multilingual models such as Whisper (Radford et al., 2022), XLS-R (Babu et al., 2021), and Omnilingual ASR (Omnilingual ASR Team et al., 2025), have achieved strong performance on high-resource standard languages. However, these improvements do not readily extend to low-resource or dialectally diverse languages, where recognition accuracy degrades substantially (Blaschke et al., 2025). Systematic evaluation on endangered languages is essential to understand model behaviour and guide preservation strategies.

This paper addresses the intersection of two critical needs: (i) the development of a systematic framework for digital preservation of Aromanian using knowledge management (KM) principles (Pendevska, 2019), and (ii) the first empirical evaluation of ASR technology on Aromanian speech. We propose a three-module framework encompassing localization, distribution, and creation of linguistic resources, and within this framework evaluate Whisper models on two Aromanian varieties—Gramosteanj and Crushova—across multiple language settings. Our contributions are fourfold: we establish the first ASR benchmark for Aromanian; we conduct a systematic evaluation across model sizes and language configurations; we analyze the impact of script mismatch and romanization; and we situate these findings within a replicable KM framework for endangered language communities, validated through empirical analysis of knowledge management practices.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 327–336
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

The significance of this research extends beyond the Aromanian language. The proposed framework offers a replicable model for endangered language preservation that can be adapted for similar linguistic communities worldwide. By integrating traditional linguistic documentation methods with modern digital technologies and KM practices, this study contributes to both language preservation efforts and knowledge management theory.

2. Background and Theoretical Framework

2.1. The Aromanian Language and Its Status

Aromanian (also known as Vlach or Macedo-Romanian; ISO 639-3: rup) is an Eastern Romance language spoken in fragmented communities across the southern Balkans, with the largest concentrations in Greece, North Macedonia, Albania, and Romania. Due to prolonged multilingual contact, Aromanian has absorbed substantial lexical, phonological, and morphosyntactic influence from Greek, Albanian, and South Slavic languages (Capidan, 1932; Friedman, 2012). An educational programme exists in North Macedonia, where the language has been taught as an elective subject in grades 3–9 since 2007.

Early scholarly documentation by Weigand (1895) and Capidan (1932) established the foundational linguistic record for Aromanian, distinguishing northern dialects (Farsheroti and other Albanian groups) from southern dialects (Gramosteanj and groups around Mount Olympus), with the Farsheroti representing a distinct variety. Marković (2007) further refined this classification, identifying that northern dialects (Farsheroti, Moskopoleans, and Muzachins) lack the characters ã and â found in southern dialects (Gramosteanj and Pindeani). Contemporary scholars including Nastev (1988) and Kahl (2008) have continued this tradition in areal linguistic studies and comparative analyses. Figure 1 illustrates the geographic distribution of the Eastern Romance language varieties in the southwestern Balkans, highlighting the fragmented nature of Aromanian-speaking communities.

2.2. Standardization and Orthographic Challenges

The systematization of Aromanian writing began approximately one century ago with the Romanian schools movement in Macedonia. The initial orthographic approach adopted the Romanian alphabet but faced challenges in representing phonemes unique to Aromanian. The language remained unstandardized until the 1997 Bitola Standardization

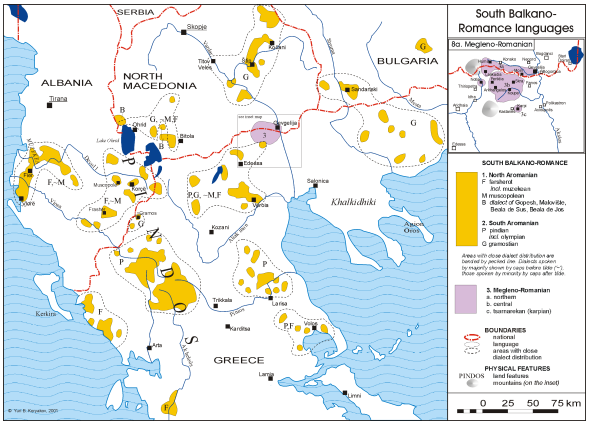


Figure 1: Distribution and dialects of the Eastern Romance languages in the southwestern Balkans. Source: Wikimedia Commons (CC BY-SA 3.0).

Symposium, which established foundational conventions by adopting a 35-character alphabet designed for compatibility with digital systems (Cunia, 2000). This represents a compromise between phonemic accuracy and technological accessibility, accommodating dialectal variation while maintaining orthographic unity (Marković, 2007).

However, the orthographic question remains sensitive and unresolved across the broader Aromanian-speaking world. The Bitola standard has been adopted by most Aromanian writers in North Macedonia, Serbia, Albania, Bulgaria, and Romania, but it is not officially recognized in Greece, where Aromanian has no formal status and alternative orthographic approaches have been proposed. Notably, Beis and Dasoulas (2017) present a Latin-based writing system developed through the teaching activities of the Society of Aromanians of Athens, designed to capture Aromanian sounds that do not exist in Greek or Romanian. Their proposal illustrates that even recent approaches from the Greek side converge on Latin-based orthography, while diverging from the Bitola conventions in specific character choices. This ongoing debate reflects the broader sociolinguistic complexity of standardizing a language spoken across multiple national contexts with differing language policies.

In North Macedonia, an east–west dialectal opposition exists, divided by the course of the Vardar river. The two varieties examined in this study—Gramosteanj (originating from the Gramos mountain region, spoken dominantly in the eastern part of North Macedonia) and Crushova (from Kruševo, western part of North Macedonia)—represent geographically and linguistically distinct communities within this division. The Gramosteanj dialect belongs to the southern variety group, while the Crushova dialect is a northern variety (Marković, 2007; Nastev, 1988). Both are written in the Latin script with community-specific orthographic con-

ventions. This late and incomplete standardization, combined with urbanization and globalization pressures, has contributed to the language's current endangered status.

Contemporary Aromanian communities are dispersed globally, with speakers found across the Balkan Peninsula (North Macedonia, Greece, Albania, Serbia, Romania) and diaspora communities in the USA and Australia. This geographic dispersion presents challenges but also opportunities for digital preservation, as digital platforms can connect speakers beyond traditional boundaries while also capturing dialectal variations.

2.3. Knowledge Management Theory for Language Preservation

Knowledge management theory provides a structured approach to linguistic preservation extending beyond traditional archival methods. The Knowledge Management Cycle (KMC) model (Evans et al., 2014) offers a seven-phase framework—identify, preserve, share, use, learn, improve, and create new knowledge—that aligns with the needs of endangered language communities, which must simultaneously preserve existing knowledge and create content for contemporary use.

Pendevska (2019) demonstrated how KM principles can be systematically applied to innovation processes, examining knowledge management practices at the individual, group, and organizational levels. This work provides the theoretical foundation for the framework proposed in this study, establishing that systematic approaches to knowledge flow organization are essential for effective preservation and distribution of knowledge. Mittelmann et al. (2022) further emphasize that effective KM requires systematic organization and distribution networks ensuring knowledge reaches relevant stakeholders with appropriate quality. For endangered languages, this means creating platforms that serve both community members wishing to maintain cultural identity and external stakeholders interested in studying linguistic diversity.

The relentless advance of digital technologies has transformed the possibilities for language study. Text-to-speech synthesis, automated language processing, and online learning platforms now enable endangered languages to participate in digital communication ecosystems (Rehm and Uszkoreit, 2012). The development of speech corpora and linguistic resources for low-resource languages has become increasingly feasible through advances in machine learning and natural language processing (Besacier et al., 2014). Digital technologies offer opportunities for creating interactive learning environments and preserving linguistic and cultural knowledge (Bird and Simons, 2002), while advances in ASR and NLP create new possibilities for endangered language participation in digital ecosystems (Joshi et al., 2020).

2.4. ASR for Endangered and Dialectal Speech

ASR for dialectal and non-standard speech remains a persistent challenge. Performance degradation has been documented for regional varieties across diverse language families, including Arabic (Nasr et al., 2023), Japanese (Takahashi et al., 2024), German (Blaschke et al., 2025; Sicard et al., 2023), and Indian languages (Alumäe et al., 2023).

Particularly relevant is recent work on Modern Greek dialects. Vakirtzian et al. (2025) introduced the first ASR benchmark for Greek varieties and showed that zero-shot Whisper performance degrades sharply with dialectal distance from Standard Modern Greek (SMG), with WER exceeding 100% for contact-influenced varieties. Tsoukala et al. (2026) expanded this benchmark, confirming a strong interaction between dialectal distance and ASR difficulty: southern dialects achieved WERs of 60–70%, while the contact-influenced Cappadocian yielded WER near 97%. Fine-tuning substantially reduced error rates across all varieties. These findings are directly relevant to Aromanian, which as a non-Greek Romance language with extensive Balkan contact features and no representation in ASR training data, presents an extreme out-of-distribution case.

3. Knowledge Management Framework for Aromanian

3.1. Three-Module Framework

We propose a three-module framework grounded in KM principles for the systematic preservation and revitalization of Aromanian (Pendevska, 2019). The framework is organized around the KMC model (Evans et al., 2014), with each module addressing specific functional requirements while maintaining integration across the full knowledge management lifecycle:

Module 1: Language Localization focuses on identifying, collecting, and digitally preserving existing linguistic and cultural resources. This encompasses historical documentation (Weigand, 1895; Capidan, 1932), television archives spanning nearly 30 years of Aromanian-language broadcasting in North Macedonia, and family sources. Content collection from family archives employs structured questionnaires covering personal data, family background, Aromanian language proficiency and identity status, and cultural practices. The dialectal mapping methodology includes audio record-

ings across different speaker communities, creating geographically representative documentation. Additionally, community engagement in populating the Mozilla Common Voice database (Mozilla Foundation, 2025) provides the foundation for building a contemporary speech corpus. Technical requirements include multiple file formats (television archives, personal recordings, visual documentation, slides, PDF) with intuitive navigation suitable for older users with limited digital experience.

Module 2: Language Learning and Distribution enables creation of accessible platforms for language education and community engagement. The digital platform provides tutorial lessons using synchronous and asynchronous media, incorporating active Aromanian teachers who contribute their pedagogical expertise for broader distribution. A key achievement is the Aromanian riddle generator presenting 100 riddles from Prof. Dr. Nastev's collection (Nastev, 1988) in Aromanian, Macedonian, and English (Pendevska et al., 2024), representing the first digital educational tool for Aromanian language learning. Distribution strategies emphasize community engagement through user-generated content creation, with integration of radio broadcasting, podcasts, and YouTube channels providing Aromanian content for wider audiences.

Module 3: Language Development and New Knowledge Creation addresses technological innovation. A significant milestone was the integration of Aromanian into the Mozilla Common Voice platform in May 2025, marking the first inclusion of an endangered and unstandardized language in this major open-source speech collection initiative (Mozilla Foundation, 2025). Speech synthesis development leverages collected data to create Aromanian-specific text-to-speech voices, facilitating language learning among younger digital-native generations and enabling integration of augmentative and alternative communication (AAC) devices worldwide. Language corpus processing enables automated linguistic analysis and supports machine learning applications for language learning and translation. Within this module, we evaluate the capabilities and limitations of current ASR technology for Aromanian, as reported in the following sections.

3.2. Empirical Validation Through KM Survey

Primary data collection included a survey examining KM practices among 176 respondents, conducted in 2017 (Pendevska, 2019). The research instrument addressed information gathering behaviours, evaluation criteria, and sharing practices at the individual, group, and organizational levels. The empirical analysis informs and validates the framework design through the findings reported below.

Information source	**%**
Internet / search engines / multimedia	27.1
Internal documents	18.8
Colleagues / experts	17.6
Training / seminars	14.1
Other sources	22.4

Table 1: Information sources used by respondents (N=176), adapted from Pendevska (2019).

Evaluation criterion	**%**
Information quality	28.3
Personal utility	24.1
Source reliability	19.3
Timeliness	15.2
Other criteria	13.1

Table 2: Most valued criteria for evaluating information (N=176), adapted from Pendevska (2019).

Table 1 presents results on information-seeking behaviour. Internet and search engine use (27.1%) is the dominant channel, followed by internal documents (18.8%), indicating strong readiness for digital language platforms. This finding suggests that digital approaches to language learning will align with established user behaviours.

Table 2 reports information evaluation criteria. Information quality (28.3%) and personal utility (24.1%) emerge as the most valued criteria, supporting the framework's emphasis on high-quality, relevant digital content rather than simply digitizing existing materials without curation.

The survey reveals strong efficiency expectations: a majority of respondents (65.5%) expect to find personal information within five minutes, while shared organizational information requires somewhat longer (46.7% find it within 5–15 minutes). These expectations indicate that digital language platforms must prioritize intuitive navigation and efficient search capabilities.

Table 3 presents information sharing practices. Text remains the primary format (29.6%), followed by tabular data (26.2%) and images/presentations (22.6%), while multimedia recordings are less common (7.5%), suggesting potential for growth in audiovisual language content. E-mail is the most used sharing channel (34.9%), followed by meetings (18.1%).

These findings validate each module: the prevalence of digital information seeking (27.1%) and internal document consultation (18.8%) support Module 1's focus on searchable digital archives; the preference for e-mail communication (34.9%) and meeting-based sharing (18.1%) validate Module 2's

Sharing format	%
Text	29.6
Tabular data	26.2
Images / presentations	22.6
Multimedia recordings	7.5
Other formats	14.1

Table 3: Information sharing formats used by respondents (N=176), adapted from Pendevska (2019).

Dialect	Original	Processed
Crushova	55:42	39:45
Gramosteanj	42:17	42:14

Table 4: Audio duration (mm:ss) before and after preprocessing.

provision of both asynchronous and synchronous learning opportunities; and the prioritization of information quality (28.3%) and personal utility (24.1%) support Module 3's focus on creating relevant, high-quality technological tools.

4. ASR Evaluation

4.1. Speech Data

Speech data for the ASR evaluation were collected from native speakers of the Gramosteanj and Crushova varieties reading dialectal texts written following the orthographic conventions discussed in Section 2.2. For standardization, all audio recordings were converted to single-channel (mono) WAV files with a sampling rate of 16 kHz.

Initially, the audio and corresponding texts were automatically aligned using the CTC Forced Aligner (CTC Forced Aligner Team, 2024). Following this automated step, the alignment boundaries were manually inspected and corrected in Praat (Boersma and Weenink, 2024) to ensure accuracy. The resulting TextGrids and audio files were then used to segment the recordings into individual utterance-level files for the evaluation phase. During this process, non-speech material such as long pauses and other untranscribed portions was removed to isolate clean speech segments. As a result, the effective audio duration was reduced for some datasets, as reflected in Table 4. Ground-truth transcriptions follow the orthographic conventions of each speaker community, reflecting the Bitola-based Latin-script standard with community-specific adaptations.

The evaluation data, including audio segments and aligned transcriptions, are publicly available on OSF. Publishing this data as the first open ASR evaluation set for Aromanian supports reproducibility and enables future benchmarking by the research community.

4.2. Models and Language Settings

We evaluate three Whisper models (Radford et al., 2022): medium, large-v2 (trained on ∼680k hours), and large-v3 (trained on ∼1M hours of weakly labeled audio plus ∼4M hours of pseudo-labeled data). All evaluations are zero-shot, without Aromanian-specific adaptation.

Since Aromanian is not supported by Whisper, we test multiple target language settings selected for linguistic proximity and contact history: Albanian (sq), Latin (la), Macedonian (mk), Greek (el), Bulgarian (bg), Romanian (ro), and automatic detection. These settings were chosen to cover genealogically related languages (Latin, Romanian), contact languages (Greek, Albanian, Macedonian, Bulgarian), and the model's own language identification capability. Romanian and Bulgarian were additionally motivated by the observation that automatic language detection frequently classified Aromanian utterances as Romanian or Bulgarian. The same set of models and language settings was evaluated for both varieties.

4.3. Evaluation Metrics and Romanization

Performance is measured using two standard metrics: Word Error Rate (WER) and Character Error Rate (CER). WER compares the machine-generated transcription to the ground-truth reference word by word, calculated as the total number of word substitutions, deletions, and insertions divided by the total number of reference words. CER operates at the character level, making it particularly useful for morphologically rich languages and for evaluating fine-grained accuracy. In cases where the target or detected language utilizes a non-Latin script (e.g., Greek, Macedonian, Bulgarian), the model's output was transliterated to the Latin script using the universal romanizer uroman (Hermjakob et al., 2018), ensuring a phonetically grounded comparison between the model's transcription and the Latin-script ground truth.

5. Results

5.1. Gramosteanj

Table 5 reports results for the Gramosteanj dialect. The best overall performance is achieved by large-v3 with Latin, yielding the lowest WER (91.60%) and CER (34.45%)—the only setting with WER below 100%.

Albanian consistently produces the lowest CER among non-Latin-output settings (42–44%), sug-

gesting phonological overlap between Aromanian and Albanian is partially captured at the character level. Romanian yields competitive results, with large-v3 achieving WER of 99.30% and CER of 43.24% (39.27% after romanization), notable given the genealogical proximity between the two languages. For Cyrillic-output languages, standard CER exceeds 90% due to script mismatch; after romanization, CER drops substantially (e.g., Macedonian from 90.78% to 41.90% and Bulgarian from 90.72% to 46.17% with large-v3).

5.2. Crushova

Table 6 reports results for the Crushova dialect. The pattern mirrors Gramosteanj: **large-v3 with Latin** achieves the best performance (WER 88.67%, CER 35.57%)—again the only sub-100% WER configuration.

Romanian yields competitive CER with large-v3 (49.26%), notable given that Romanian is genealogically close major language. However, the model's Romanian decoder still cannot reliably reconstruct Aromanian words (WER 100.99%). Bulgarian and Macedonian show similar CER patterns, with standard CER above 90% dropping to approximately 45–54% after romanization.

5.3. Cross-Dialect Comparison

The two dialects exhibit remarkably similar performance profiles. Crushova achieves slightly lower best-case WER (88.67% vs. 91.60%) and comparable best-case CER (35.57% vs. 34.45%). This consistency across two geographically and sociolinguistically distinct communities suggests that the observed performance levels reflect fundamental model limitations with respect to Aromanian rather than dialect-specific properties.

6. Discussion

6.1. Comparison with Greek Dialect ASR

Our results can be contextualized by comparison with Greek dialectal ASR benchmarks. Vakirtzian et al. (2025) reported zero-shot Whisper large-v3 WER of 58.42% on Eastern Cretan and 32.71% on Messenian—varieties close to SMG—but WER exceeding 100% for contact-influenced Aivaliot and Griko. Tsoukala et al. (2026) showed WERs of 61.85% (Aperathiot), 70.24% (Cretan), 80.87% (Lesbian), and 96.65% (Cappadocian).

Aromanian, with best zero-shot WER of 88–92%, falls between Lesbian and Cappadocian in difficulty. This is consistent with Aromanian's position as a language entirely unrepresented in training data and exhibiting substantial contact-induced lexical borrowing. The parallel with Cappadocian—both are contact-influenced varieties representing extreme out-of-distribution cases—is particularly instructive.

6.2. Language Setting and Script Effects

The Latin setting consistently outperforms all other configurations, likely because it eliminates script mismatch and shares phonological features with Aromanian as a fellow Romance language. Albanian, despite producing Latin-script output, yields higher WER, suggesting lexical and morphological differences outweigh script compatibility. The romanization experiments demonstrate that a large portion of CER for Cyrillic-output languages reflects script mismatch: after romanization via uroman, Macedonian CER drops by approximately 50 percentage points, reaching levels comparable to Albanian. This parallels the Griko findings of Vakirtzian et al. (2025).

6.3. WER–CER Divergence

Across all configurations, WER substantially exceeds CER, most pronounced for the best settings (e.g., large-v3 Latin: 91.60% WER vs. 34.45% CER for Gramosteanj). This indicates that while models fail to reconstruct correct word forms, they capture meaningful phonotactic and subword structure. The lower CER for Aromanian under optimal settings (34–36%) compared to Cappadocian (59.72%) may reflect the benefit of native Latin-script alignment.

6.4. Implications for the Preservation Framework

The ASR evaluation results have direct implications for the KM framework (Pendevska, 2019). Within Module 3, the findings demonstrate that current off-the-shelf ASR models cannot produce usable Aromanian transcriptions without dedicated adaptation. This underscores the importance of community-driven data collection initiatives such as the Mozilla Common Voice integration, which will provide the training data necessary for future model fine-tuning. The high CER under optimal conditions (34–36%) suggests that fine-tuning on even modest amounts of dialectal data—following the approach shown effective for Greek dialects (Tsoukala et al., 2026)—could yield substantial improvements, potentially enabling practical ASR applications for documentation and education within the Module 2 distribution platform.

The empirical KM findings further reinforce these implications. The dominance of digital information seeking (27.1%) and the strong efficiency expectations (65.5% expect results within five minutes) indicate that digital language preservation

Model	Lang	WER (%)	CER (%)	WER rom. (%)	CER rom. (%)
medium	sq	100.80	42.15	—	—
medium	la	120.86	80.71	—	—
medium	mk	108.94	93.69	103.79	47.42
medium	el	111.14	98.85	109.59	68.44
medium	bg	109.90	92.40	104.62	46.97
medium	ro	100.78	46.16	99.90	43.02
medium	auto	106.83	71.84	103.79	45.85
large-v2	sq	109.20	43.90	—	—
large-v2	la	97.87	41.87	—	—
large-v2	mk	106.31	91.35	99.37	42.36
large-v2	el	109.67	97.96	108.03	64.39
large-v2	bg	109.43	92.47	103.46	46.00
large-v2	ro	102.36	45.70	100.94	42.10
large-v2	auto	105.35	58.87	103.31	43.92
large-v3	sq	112.09	44.26	—	—
large-v3	la	**91.60**	**34.45**	—	—
large-v3	mk	106.15	90.78	100.36	41.90
large-v3	el	106.45	94.36	104.73	60.11
large-v3	bg	107.39	90.72	102.34	46.17
large-v3	ro	99.30	43.24	97.72	39.27
large-v3	auto	107.32	68.98	104.19	44.83

Table 5: ASR results on the Gramosteanj dialect. Romanized (rom.) metrics are reported where transliteration to Latin script was applied. Best results in bold.

Model	Lang	WER (%)	CER (%)	WER rom. (%)	CER rom. (%)
medium	sq	101.78	45.83	—	—
medium	la	105.79	72.83	—	—
medium	mk	110.25	93.55	106.35	54.05
medium	el	113.29	100.83	112.18	82.25
medium	bg	111.26	93.25	107.11	54.50
medium	ro	114.12	63.02	113.02	60.31
medium	auto	104.73	62.02	102.32	47.00
large-v2	sq	108.57	46.44	—	—
large-v2	la	97.14	42.77	—	—
large-v2	mk	109.15	92.08	104.24	47.92
large-v2	el	113.14	101.35	112.13	77.34
large-v2	bg	111.26	92.84	106.77	51.98
large-v2	ro	109.69	58.39	108.39	55.26
large-v2	auto	108.40	57.82	107.24	50.00
large-v3	sq	112.43	46.48	—	—
large-v3	la	**88.67**	**35.57**	—	—
large-v3	mk	108.61	91.07	103.45	45.42
large-v3	el	110.29	97.17	109.39	74.05
large-v3	bg	110.62	91.25	106.35	54.04
large-v3	ro	100.99	49.26	99.55	46.22
large-v3	auto	107.74	63.12	105.27	47.97

Table 6: ASR results on the Crushova dialect. Romanized (rom.) metrics are reported where transliteration to Latin script was applied. Best results in bold.

platforms must prioritize intuitive navigation and efficient search capabilities. The emphasis on information quality (28.3%) as the top evaluation criterion confirms that automated transcription tools, once sufficiently accurate, would be highly valued by the target user community. The relatively low current usage of multimedia recordings for sharing (7.5%) suggests significant potential for growth as speech technologies improve.

7. Conclusion

This paper presents an integrated approach to Aromanian language preservation, combining a knowledge management framework with the first systematic ASR evaluation for this endangered language. The three-module KM framework—encompassing localization, distribution, and creation (Pendevska, 2019)—provides a replicable model for endangered language communities, validated through empirical analysis of KM practices (N=176) and supported by concrete achievements including Mozilla Common Voice integration and the first digital Aromanian riddle generator (Pendevska et al., 2024).

The ASR evaluation demonstrates that current multilingual models fail to produce acceptable Aromanian transcriptions in zero-shot settings (WER $>$ 88%), but CER as low as 34–36% under optimal conditions indicate meaningful subword capture. The Latin language setting with Whisper large-v3 consistently achieves the best results, and romanization of non-Latin output substantially improves CER. These findings position Aromanian alongside contact-influenced Greek varieties as one of the most challenging cases for pretrained multilingual models.

The empirical KM findings—particularly the dominance of digital information seeking, the prioritization of information quality, and the strong efficiency expectations—provide actionable design guidelines for digital language preservation platforms. Future work should explore fine-tuning on community-collected data, evaluate additional model architectures, and conduct longitudinal studies tracking the framework's effectiveness in achieving language revitalization. The Aromanian case demonstrates that with systematic KM approaches and realistic understanding of technological capabilities, endangered languages can begin transitioning from passive documentation to active participation in digital ecosystems.

8. Limitations

All ASR evaluations are zero-shot; fine-tuning experiments could not be performed due to limited data and the absence of standard train/dev/test splits. The absence of a standardized Aromanian orthography means ground-truth transcriptions reflect individual transcriber conventions, potentially inflating error rates. Only Whisper models were evaluated; other architectures (XLS-R, Omnilingual ASR) may yield complementary insights. The aggregate WER and CER metrics do not reveal which specific linguistic phenomena contribute most to errors; qualitative error analysis is left for future work. Additionally, the survey data informing the KM framework was collected in 2017 and examines general KM practices rather than language-specific behaviours, potentially limiting direct applicability to the language preservation context. The geographic and demographic scope of the survey may not fully represent global Aromanian communities, particularly diaspora communities with different technological access and cultural preservation priorities.

9. Data Availability

The oral speech data used for the ASR evaluation in this study are publicly available via the Open Science Framework (OSF) repository at the following link: `https://osf.io/z2n7r/overview`.

The repository contains the processed audio segments and corresponding aligned transcriptions used for evaluation, ensuring full reproducibility of the reported results

Ethics Statement

All speech data were collected with informed consent. The survey was conducted with voluntary participation. The goal is to support linguistic documentation and technological inclusion, not deployment in real-world applications. The reported ASR models exhibit high error rates and should not be used for practical applications involving Aromanian speakers.

Acknowledgments

First and foremost, I am humbled and give my deepest respect to Ms. Ljubica Georgievska for her dedicated work on keeping the Aromanian language alive and contemporary.

10. Bibliographical References

Tanel Alumäe, Jiaming Kong, and Daniil Robnikov. 2023. Dialect Adaptation and Data Augmentation for Low-Resource ASR: TalTech Systems for the MADASR 2023 Challenge.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

Stamatis Beis and Fanis Dasoulas. 2017. Proposal for the writing system of the Vlach language. *Glossologia*, 25:51–69.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Steven Bird and Gary Simons. 2002. Seven dimensions of portability for language documentation and description.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, pages 913–917. ISCA.

Theodor Capidan. 1932. *Aromânii: Dialectul aromân*. Imprimeria Naţională, Bucharest.

David Crystal. 2000. *Language Death*. Cambridge University Press.

Tiberius Cunia. 2000. On the standardization of the Aromanian system of writing. *The Farsarotul*.

M. Evans, K. Dalkir, and C. Bidian. 2014. A holistic view of the knowledge life cycle: The knowledge management cycle (KMC) model. *Electronic Journal of Knowledge Management*, 12(2):85–97.

Victor A. Friedman. 2012. Aromanian in the Balkan linguistic league: A comparative approach. volume 14, pages 51–66.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Thede Kahl. 2008. Does the Aromanian have a chance of survival? Some thoughts about the loss of language and language preservation. *The Romance Balkans*, pages 123–140.

Marjan Marković. 2007. *Govorot na Aromancite Frašeroti vo Ohridsko-Struškiot region (vo Balkanski kontekst)*. MANU, Skopje. Doctoral dissertation defended in 2000.

Angelika Mittelmann, Gabriele Vollmar, et al. 2022. Wissensmanagement-kompetenzkatalog. Version 2.0. Gesellschaft für Wissensmanagement e.V. (GfWM).

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Seham Nasr, Rehab Duwairi, and Muhannad Quwaider. 2023. End-to-end speech recognition for arabic dialects. *Arabian Journal for Science and Engineering*, 48.

Božidar Nastev. 1988. *Aromanski studii, prilozi kon balkanistikata*. MANU, Skopje.

Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages.

Marija Pendevska. 2019. *The Impact of Knowledge Management on Innovation in Enterprises in the Republic of Macedonia*. Ph.D. thesis, Faculty of Economics, Ss. Cyril and Methodius University in Skopje, Skopje.

Marija Pendevska, Branislav Gerazov, and Branko Prlja. 2024. Digital preservation of Aromanian through a knowledge management framework: The first digital riddle generator. In *Proceedings of the 2nd International Conference Dedicated to 100 Years Prof. Dr. Božidar Nastev*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Georg Rehm and Hans Uszkoreit. 2012. Language technology 2012: Current state and opportunities. In *The META-NET Strategic Research Agenda for Multilingual Europe 2020*, pages 27–31.

Clement Sicard, Kajetan Pyszkowski, and Victor Gillioz. 2023. Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects.

Naoki Takahashi, Shogo Miwa, Yuta Kamiya, Takumi Toyama, Raufun Nahar, and Atsuhiko Kai. 2024. Comparison of Large Pre-trained Models and Adaptation Methods for Japanese Dialects ASR. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 811–814.

Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet

Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos, and Stella Markantonatou. 2026. Extending ASR evaluation resources for Modern Greek dialects. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 210–222, Rabat, Morocco. Association for Computational Linguistics.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Gustav Weigand. 1895. *Die Aromunen: Ethnographisch-Philologisch-Historische Untersuchung über das Volk der sogenannten Makedo-Romanen oder Zinzaren*. J.A. Barth, Leipzig.

11. Language Resource References

Boersma, Paul and Weenink, David. 2024. *Praat: Doing Phonetics by Computer*. University of Amsterdam. PID https://www.praat.org/.

CTC Forced Aligner Team. 2024. *CTC Forced Aligner*. Open Source. PID https://github.com/huggingface/ctc-forced-aligner.

Hermjakob, Ulf and May, Jonathan and Knight, Kevin. 2018. *Out-of-the-Box Universal Romanization Tool Uroman*. USC Information Sciences Institute. PID https://github.com/isi-nlp/uroman.

Mozilla Foundation. 2025. *Mozilla Common Voice*. Mozilla Foundation. PID https://commonvoice.mozilla.org/.

A Novel Typology of Mutually Intelligible Words: The Case of Slavic Languages

Klyshinsky E., Badryzlova Yu.
HSE University
Staraya Basmannaya Str. 21/4 Bld. 1, Moscow, Russia
yuliya.badryzlova@gmail.com, eklyshinsky@hse.ru

Abstract

In this paper, we demonstrate that using the notion of cognates in the task of evaluating mutual intelligibility (MI) in closely related languages can be confusing. We suggest a new term – percipiants – which handles MI of words irrespective of whether they have a common origin. We propose four classes of percipiants,which are differentiated by the degree and type of the closeness of meanings within the word pair. Furthermore, we claim that MI of individual words across a set of languages may be established computationally via normalized Levenshtein Distance (LD). We test our hypotheses by analyzing data from a psycholinguistic experiment in which the respondents had to predict missing words in a text in their native language. In the experimental group, the respondents had access to the text in their native language and in a language from the same language group (Slavic); in the control condition, the respondents were performing the same test in the presence of their native language only. The analysis demonstrates that (a) psycholinguistically, MI may be defined as the difference between the average correctness of answers in the experimental and the control groups; (b) normalized LD may serve as an adequate predictor of the experimentally measured MI; (c) the said MI corroborates the four classes of percipiants; (d) contextual factors weaken the predictive force of LD, requiring further investigation.

Keywords: mutually intelligible words, cognates, percipiants

1. Introduction

The task of assessing mutual intelligibility (MI) of texts written in different genetically related languages has attracted heightened interest in foreign language instruction (e.g. Squires et al. 2020), as well as in studies of bilingualism (e.g. Miwa et al. 2014) and multilingualism (e.g. Stenger et al. 2020). Intelligibility of a foreign language is understood as a situation in which speakers of one language are capable of understanding written or spoken texts in another language without prior training or preparation. Typically, texts in closely related languages are intuitively comprehensible even to speakers with no professional linguistic background, while the intelligibility of distant languages is known to depend on the presence and scope of common borrowings from a third language (or languages).

It is commonly accepted that MI is promoted primarily via cognates. Consequently, a substantial body of research has focused on the impact of cognates bearing orthographic (Gooskens and Heeringa, 2004; Van Bezooijen and Gooskens, 2005) as well as phonetic (Gooskens and Heeringa, 2004; Stenger et al., 2020) similarity. Yet, empirically-based intuition prompts that cognates are not the only factor behind MI. On the one hand, some cognates may be poorly recognized by the recipient: this occurs when a common cognate base attaches language-specific affixes or undergoes substitutions of letters or sounds; when a word is written in a different alphabet, or when the orthographic characters have language-specific diacritics. On the other hand, meaning may be reliably inferred not only for cognates. Namely, MI is possible for words which are not true cognates, yet they possess phonetic or graphic similarity and a shared meaning – though such cases are quite rare, cf. *much* (English) vs. *mucho* (Spanish). (Here, we are setting aside the discussion of the distinction between cognates and borrowings, which is beyond the scope of this paper.)

Tracing etymological roots is typically very challenging, which, in turn, complicates the assessment of MI and impedes its practical application (e.g., in language teaching). Consequently, we consider it reasonable for practical purposes to distinguish between the shared ancestry of words and their MI.

We introduce a new term — *percipiant* — to denote words that are mutually intelligible across languages irrespective of whether they share a common origin; this allows us to shift the focus from similarity of origin toward similarity of comprehension. In order to gain insight into mutual cross-linguistic intelligibility, we analyze the results of an experiment in which native speakers of Russian were to fill in missing words in a Russian text while also being exposed to a parallel translation into one of eight Slavic languages. The MI of a word was computed as the increase in correct answers in the presence vs. absence of parallel texts.

Our results show that percipiants can be divided

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 337–344
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

into four groups based on their similarity, as manifested in the different statistical patterns of respondents' behavior in recovering the missing words. We also demonstrate a correlation between Levenshtein distance (Levenshtein, 1966) and MI; however, we conclude that Levenshtein distance alone does not fully account for the variation in this behavior; contextual factors exert a strong influence as well, thus calling for a closer analysis.

2. Previous work *vs.* need for a novel approach

Research on MI in genetically related languages has largely focused on analysis of cognates. Crystal (2008, p. 83) defines cognates as follows: "A language or a linguistic form which is historically derived from the same source as another language/form, e.g. Spanish/Italian/French/Portuguese are 'cognate languages' (or 'cognates'); *père*/*padre*, etc. ('father') are 'cognate words' or cognates" – thereby stressing the common ancestry. The definition by Squires et al. (2020) states that cognates are "...pairs of vocabulary words that share meaning with similar phonology and/or orthography in more than one language, such as *rose–rosa* (English–Spanish) or *carrot-carotte* (English–French)". According to Miwa et al. (2014), "The words *coffee* /khɔfi/, *koffie* /kɔfi/, and コーヒー /koohii/ ... are examples of cognates" – thus stressing the role of phonetic resemblance for the MI of cognates.

Roembke et al. (2024) conducted a meta-analysis of a large body of studies on MI. They found that "Operationalizations of cognate status varied greatly across studies, though three methods were most common: (1) similarity was quantified by using some kind of norming procedure to obtain similarity ratings (i.e., people either classified word pairs as cognate/non-cognate or rated them on a similarity scale…), (2) (normalized) Levenshtein distance,… or (3) Van Orden's graphemic similarity algorithm (Van Orden, 1987)". Method 1 mentioned in the meta-analysis indicates that it is not uncommon in research practice to establish cognates via the perceived intelligibility judged by human subjects, whereas common historical origins are not taken into account.

Levenshtein distance (LD) is one of the most commonly used metrics for measuring the orthographic similarity of words. MI is often analyzed via its correlation with LD (e.g. Van Bezooijen and Gooskens 2005). Thus, LD was shown to adequately quantify orthography-based similarity across entire languages (Golubović and Gooskens, 2015; Gooskens and Heeringa, 2004). At the same time, attempts to apply apply LD to individual words demonstrated a heavily diffuse pattern, with normalized LD values ranging between 0.0 and 0.5 (Saturno, 2023; Stenger et al., 2020).

Roembke et al. (2024) cite works where words are recognized as cognates at LD > 0.83 and as non-cognates at LD < 0.14. This approach, again, treats cognates as words being similar rather than having common historical roots. A notable omission of this approach is that it fails to account for the fact that inflectional factors (e.g., prefixation and suffixation) may dramatically impact graphic similarity between words – which is particularly relevant for morphologically rich languages, such as Germanic and Slavic languages.

In addition to LD, word adaptation surprisal (Stenger et al., 2020) and conditional entropy (Gooskens et al., 2007) were proposed to measure similarity of word pairs. These methods have the advantage of being asymmetrical, i.e. speakers of different languages are likely to exhibit varying intelligibility of words within the same pair of texts.

Thus, interpretation of cognates as similarly perceived words persists across a large number of studies, thereby violating the fundamental definition of cognates and leading to grave confusions. (Having said that, one should not deny the value of similarity-based methods for automated identification of candidates for true cognates.)

In order to divorce the two problems – finding common etymology and assessing MI in word pairs – the present paper introduces the notion of *percipiants* and their four classes as follows:

A) *True percipiants* are pairs of words which are perceived as being similar, but not necessarily sharing a common origin, e.g. povernulsja[1] :: povernuvsa (RU[2] повернулся :: UA повернувся turn-PST.M.SG-REFL '(he) turned (himself) around'[3]). Percipiants may be close to full, partial, or false cognates; however, we do not attempt to verify their common origin.

B) *Percipiant synonyms* are two (or more) words where a word has a synonym or a near-synonym within the same language; the (near-)synonym, in turn, is a true percipiant to the foreign counterpart of the pair, e.g. dorogu :: put :: putja (RU дорогу :: RS пут road-ACC.SG :: BG пътя road-DEF. The Russian synonym of dorogu is puti (пути road-

[1] Throughout this paper, both Latinate and Cyrilic scripts are transliterated using the normalized, pseudo-phonetic Latinization scheme (see Section 3) in order to ensure consistency of graphic representation.

[2] The following language codes are used throughout this paper: BG – Bulgarian; BY – Belarusian; CZ – Czech; PL – Polish; RS – Serbian; RU – Russian; SI – Slovene; SK – Slovak; UA – Ukrainian.

[3] In cases when a gloss and translation are valid for two or more languages (like here), the former are presented only once.

ACC.SG)).

C) Consequently, *non-percipiants* are pairs of words which are not perceived as being similar while having a common scope of meaning. This may occur due to the following:
   - they are true non-cognates, e.g. truboju :: surmoju :: polnitsou (RU трубою :: UA сурмою trumpet-INSTR.SG '(with) a trumpet' :: poľnicou SK (hunting).horn-INSTR.SG '(with) a horn').
   - they are possible cognates, yet their mutual recognizability is rather low e.g. ostanovilis :: zaustavili (RU остановились PFV-become.steady-CAUS-PST-PL-REFL 'they stopped' :: RS зауставили PFV-PREV-set/place-CAUS-PST.PL '(they) stopped (something)').

D) Finally, *false percipiants* are pairs of words in which the foreign counterpart may be similar to the native word in its appearance, while having a different meaning, cf. spinoju :: grba (RU спиною back-INSTR.SG 'with (one's) back' :: BG гърба back-DEF 'the back'), where the Bulgarian word resembles the Russian gorba (горба hump-GEN.SG '(of one's) hump'). Similarly, povernulsja :: zavjarnuwsja (RU повернулся :: BY PFV-turn-SUF-PST.M.SG-REFL '(he) turned (himself) around'), where the Belarusian word resembles the Russian zavernulsja (завернулся PFV-wrap-SUF-PST.M.SG-REFL '(he) wrapped himself (up)'.

Bearing these four classes in mind, now let us examine the impact of (normalized) graphic proximity on mutual intelligibility of percipiants.

3. Data and methods

The data utilized in this experiment is the psycholinguistic dataset introduced in Klyshinsky and Lukashevich (2025). In the experimental group, the responses were collected from native speakers of Russian who were exposed to a Russian text concurrently with its translation counterpart in one of the eight Slavic languages: Belarusian, Bulgarian, Czech, Polish, Serbian, Slovak, Slovenian, and Ukrainian (for detail, see Klyshinsky 2021). Individual words were omitted in the Russian text; the respondents were tasked with restoring the missing words. In the control group, the same task was offered, yet with no parallel translation. In order to to standardize the representation of text for Russian-speaking respondents, the stimuli from non-Cyrillic graphic systems were presented both in the original (Latin) and the normalized Cyrillic forms, i.e. transliterations into the Cyrillic alphabet which were performed with IPA transcription (International Phonetic Association, 1999) as an intermediary.

In total, the dataset contains responses on 186 unique lemmas for 232 gaps across six texts. A total of 1,975 respondents in the experimental group and 283 respondents in the control group were surveyed. All word pairs (consisting of an omitted Russian word and its foreign counterpart) are manually tagged with one of the four classes of percipiants presented in Section 2; the number of cases per class is shown in Figure 4 (Section 4). Furthermore, the collected responses (both in the experimental and the control groups) were manually inspected and scored as follows:

- 1 point: correct answer, when a respondent chose the same word as in the original text, its synonym or near-synonym; divergencies from the original grammatical parameters were allowed.
- 0.5 points: partially correct answer, when a respondent chose a word derived from the original word yet of a different part of speech; or a word of the same part of speech which is unrelated to the original, yet without disrupting the narrative logic. The underlying assumption is that the user either understood the word while misunderstanding the syntactic structure; or they grasped the syntactic structure yet failed to find a correct lexical match.
- 0 points: wrong answer, when a respondent left the blank unfilled, or chose an entirely inadequate word – either from a wrong part of speech or disruptive to the flow of the narration.

For instance, in the phrase . . . rassuzhdal on po doroge. . . (RU . . . рассуждал он по дороге. . . '... he was reflecting as he was going down the road. . . '), the last word, road-DAT.SG is the target. Therefore, the responses (po) doroge and (po) puti (по дороге, по пути along=road-DAT.SG) are scored as correct; partially correct answers, for example, are (po) hodu (по ходу along=course-DAT.SG 'in the course (of something)'), (po) naitiju (по наитию by=intuition-DAT.SG 'intuitively'), (po) nocham (по ночам at=night-DAT.PL 'at nights'); finally, the following choices are marked as wrong: (po)-svoemu (по-своему ADVZ-one's.own-DAT 'in one's own manner'), (po-) obychnomu (по-обычному ADVZ-usual-DAT 'in the usual way'), (po-) rimski (по-римски ADVZ-Roman-NOM.SG.M 'in the Roman manner'), (po) sebe (по себе for=self-DAT 'to oneself'). It is reported that the annotation by percipient types and the scoring were performed by two native Russian linguists; the decisions of the first annotator were cross-checked by the second annotator; disagreements (~5% of the cases) were resolved by consensus.

Since the present study seeks to find a correlation between Levenshtein distance (LD) and the

correctness of respondents' answers, following Heeringa (2004) we resorted to phonetic representation of words rather than to their graphemic form. Therefore, we applied the same scheme of transcription as in the original dataset; the subsequent computations of LD on languages with Latin alphabets were performed on their normalized Cyrillic representations.The texts in languages with the Cyrillic script (other than Russian) were also normalized to ensure consistency, i.e. the characters that are not present in the Russian alphabet were substituted with their Russian equivalents (e.g., *i* → и, *j* → й).

Furthermore, in following Heeringa (2004), while computing LD we prioritized insertions and deletions over replacements, keeping equal weights across these three operations, which forces the algorithm to favor greater spans of similar characters. We used a weighted version of LD, where a transposition is assigned a total cost of 1 (i.e., 0.5 per character moved) rather than a cost of 2 (which would have been the cost for two separate substitutions). Substitutions between similar sounds (e.g., [o]-[u], [a]-[o], [ʃ]-[ʒ], [g]-[k]) were assigned a cost of 0.6.

Besides, we normalized the LD values by the length of the longest word within a pair; this occasionally resulted in the normalized LD > 1.0 for some words, that is, in the cases when transliteration increases a word's character count, for instance, when the Czech grapheme *ř* is transliterated as the respective longer Cyrillic digraph *rz*.

It is essential to note that, unlike previous studies on mutual intelligibility (Golubović and Gooskens, 2015; Gooskens and Heeringa, 2004), we analyze words as they appear in context; however, the present study does not aim to provide an in-depth analysis of the contextual factors that might impact intelligibility. This research question, therefore, remains to be explored in subsequent investigations; for present purposes, we will limit our discussion to brief observations and comments.

4. Results

As we showed above (Section 2), many studies use LD in order to measure the proximity of entire languages by averaging the values across massive collections of vocabulary. Such an approach has been shown to quite adequately capture cross-linguistic trends; thus, Gooskens & Heeringa (2004) demonstrate that mutual intelligibility (MI) is declining with the increase in LD. However, attempts to measure the MI for a pair of words (e.g. see Saturno 2023; Van Bezooijen and Gooskens 2005) yielded less impressive results. The present study also began with an attempt to straightforwardly apply normalized LD to evaluate MI of word pairs, with the MI value measured as the average score of responses from the experimental group across the prompts from all the eight foreign languages. However — quite expectedly — this attempt led to a not-so satisfactory outcome (see Figure 1).

In Figure 1, the majority of true percipiants have lower LDs as compared to the other percipient classes, with a fuzzy threshold between them running at around LD=0.5. However, some of the words marked by the experts as non-percipients occasionally have a low distance (LD < 0.2). It can also be seen from the figure that the overall trend of decreasing average scores along with increasing LDs manifests itself mainly at the lower boundary, i.e. some words will always be reliably restored irrespective of their LDs. A manual analysis of these instances indicated that, in the majority of cases, the missing word is recoverable solely from the monolingual context, irrespective of the similarity within the word pair. This happens, for example, when the missing word is a part of a frequent collocation or an idiom, e.g. in (1), where the underlined part of the sentence is a fixed collocation. Language-pairwise analysis of the data reveals that intelligibility may vary depending on the language of the non-Russian counterpart.

Interestingly, this behavior is occasionally observed even when LD is next to 0. Thus, the average score of kvadrata (see 2a) in the control group is 0.52; meanwhile, its average score in the presence of a parallel text ranges between 0.89 for UA, through 0.85 for BY and 0.76 for BG, to 0.71 for RS – although the LD to the target in all the cases is 0 (or just below): kvadrata :: kvadrata (RU, UA, BY квадрата square-GEN.SG; BG квадрата square-DEF.M.SG[GEN] 'of a square'). We presume that such variation is due to the high LD values of not the target per se, but of its surrounding context, cf. Examples 2a and 2b vs. 2a and 2c below.

Meanwhile, the responses given by the informants indicate that not all word pairs marked by the experts as true percipiants are perceived as such. Thus, the lowest average scores in the experimental group were obtained for the pairs raba :: *robski* (RU раба slave-GEN.SG :: BG робски slave-ADJ.SG 'of the slave') with the average score of 0.27 and LD=0.583; truboju :: trubkou (RU трубою :: CZ trubkou trumpet-INS.F.SG '(with) a trumpet') with the score of 0.51 and LD=0.286; and osuzhdennyh :: osudzhenike (RU осужденных condemn-PST.PTCP.PL-GEN.ACC.PL :: RS осуђенике condemn-PST.PTCP.PL-LOC.PL '(of/about) the convicted (people)') with the score of 0.3 and LD=0.5. For example, cf. (3), where the average score in the control group is 0.352 and in the experimental group it is 0.514. The poor recognizability of the target in the experimental condition is due to a number of factors, such as very low LDs

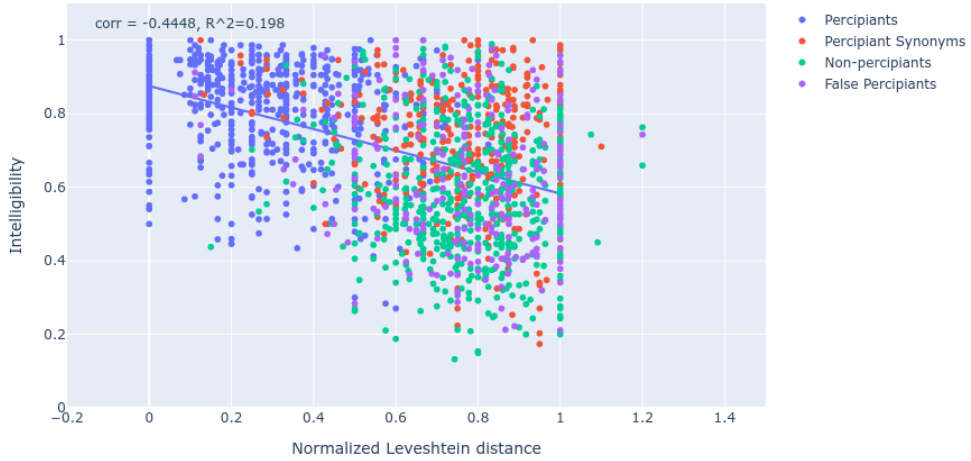


Figure 1: Distribution of mutual intelligibility among percipiants by type and LD.

(Ex. 1)

От	этой	мысли	у	Виниция	потемнело	в	глазах
Ot	etoj	mysli	u	Vinitsija	potemnelo	v	glazah
from	this.-F.GEN	thought.-GEN	at	Vinicius.-GEN	PFX-dark-PST-N.SG	at	eye-PREP.PL

'This thought made Vinicius's eyes grow dark.'

of the left and the right contexts; the profound syntactic differences which lead to differences in the linear word order (so that words from the original's left context occasionally move into the right context and vice versa) and to different syntactic means being used in a similar function (i.e. participial phrase vs. relative clause), etc. The opposite is also true, when pairs which was annotated as non-percipiants yielded high average scores, e.g. (4), where the non-percipiant pair poljubil vs. pokokhav (RU полюбил :: UA покохав PFV-love-PST.M.SG '(he) fell in love') has LD=0.688 and the average score of 0.95. Here, the high recognizability is promoted by the extra-sentential context and narrative motif which describes the protagonist's emotions and reflections caused by love. This context allows reliably restoring the missing word even when no parallel text is to help.

In contrast, consider (5) as an example of an 'ideal' percipiant: vecher :: vecher (RU вечер :: CZ večer evening-NOM.M.SG), with LD=0; the average score in the control group was 0.58, but in the experimental group it plummeted to 1.0. The reason is that the Russian sentence has a zero link verb, consequently, the gap allows for almost unrestrained combinability; in this case, the similarly-looking prompt from the parallel text provides critical assistance.

To sum up, the assessment based on word-stem similarity which is provided by experts, may not adequately align with the perception by respondents. That is, mutual intelligibility (MI) is affected by other, mainly contextual, factors except LD, as shown above.

Since the first stage of the analysis demonstrated that MI cannot be reliably quantified as a relationship between LD and the average score in the experimental group, we moved on to a more complex design. Here, MI of a word pair was computed as the difference between the average scores of the experimental and the control group (see Figure 2). The intuition behind this heuristic was that the predictability of the word in the native-language context is to increase when a parallel text in another mutually intelligible language is present. Note that MI may take both positive and negative values, i.e. presence of the parallel text may both facilitate and inhibit recognition of the target. It should be noted that if the average score for a word in the control group is close to 1, then the prompt from the parallel text will bring about little or zero increase; thus, the MI of the word pair will be close to 0.0. We realize that such cases do not adequately reflect MI and may introduce distortions into the overall result.

Therefore, they require a separate analysis and more scrupulous design of contexts in future research. Figure 2 demonstrates an increase in correlation and R^2, as compared to Figure 1 – from -0.44 to -0.53, and from 0.198 to 0.277, respectively. Most importantly, we see that true percipiants exhibit higher MI than the other classes of percipiants, alongside lower LD values. The mean MI for true percipiants is 0.208 (SD = 0.166), for percipiant synonyms it is 0.08 (SD = 0.143), for non-percipiants -0.04 (SD = 0.127), and for false percipiants -0.003 (SD = 0.138). Thus, we observe that true percipiants are, on average, significantly more intelligible than the other three classes.

Figure 3 presents the distributions of average

		1 – Avg. Score; 2 – LD, left context, 3 – LD, right context	1	2	3
(Ex. 2a)	RU	. . . на всех трех сторонах квадрата, кроме той, где были наши собеседники. . . . na vseh storonah kvadrata, krome toj, gde byli nashi sobesedniki.	0.52	NA	NA
(Ex. 2b)	UA	. . . на всіх трьох сторонах квадрата, крім тієї, де були наші бесідники. . . . na vsih troh storonah kvadrata, krim tiej, de buli nashi besedniki.	0.89	0.13	0.28
(Ex. 2c)	RS	. . . на све три стране квадрата, осим оне на којој су се налазили наши саговорници. . . . na sve tri strane kvadrata, osim one na kojoj su se nalazili nashi sagovornitsi.	0.71	0.5	0.76
		'...on all three sides of the square, except for the one where our interlocutors were.'			

(Ex. 3) RU . . . и мимо Пилата последним проскакал солдат с пылающей на солнце трубою за спиной.
. . . i mimo Pilata poslednim proskakal soldat s pylajusczcej na solntse truboju za spinoj.
CZ . . . poslední projel těsně kolem Piláta voják s trubkou na zádech, která plála v paprscích slunce.
. . . posledni projel tesne kolem Pilata vojak s trubkou na zadech, ktera plala v paprscich slunce.
'...and finally, past Pilate galloped a soldier, with a trumpet ablaze in the sunlight, slung across his back.'

(Ex. 4) RU Лишь теперь ему стало ясно, как сильно он ее полюбил.
Lish teper emu stalo jasno, kak silno on ee polubil.
UA Щойно лише йому стало ясно, як її покохав.
Szczojno lishe jomu stalo jasno, jak jj pokohav.
'Only now did he realize how much he loved her.'

(Ex. 5) RU . . . и увидел, что на Патриарших вечер.
. . . i uvidel, chto na Patriarshih vecher.
CZ . . . a viděl, že se mezitím na Patriarchovy rybníky snesl večer.
. . . a videl, zhe se mezitim na Patriarchovy rybniky snesl vecher.
'...and he saw that evening had fallen on the Patriarch's Ponds.'

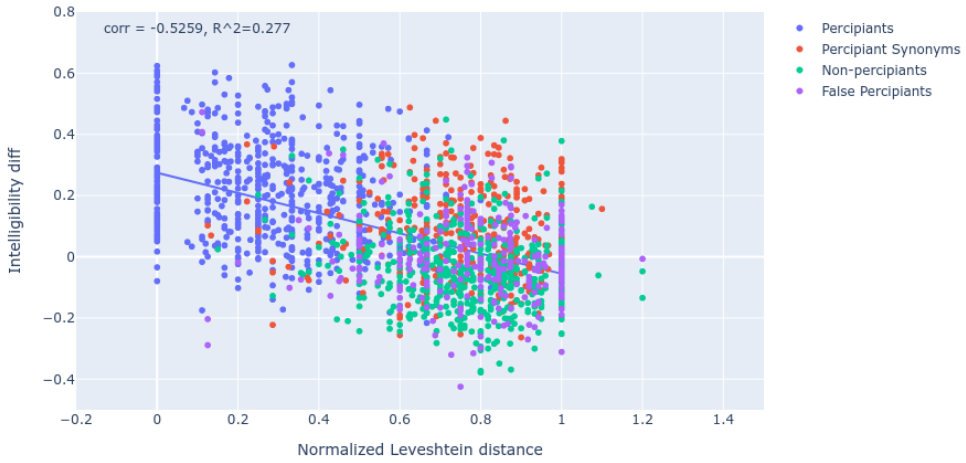


Figure 2: Distribution of the difference in intelligibility between the experimental and the control group, by the type and LD.

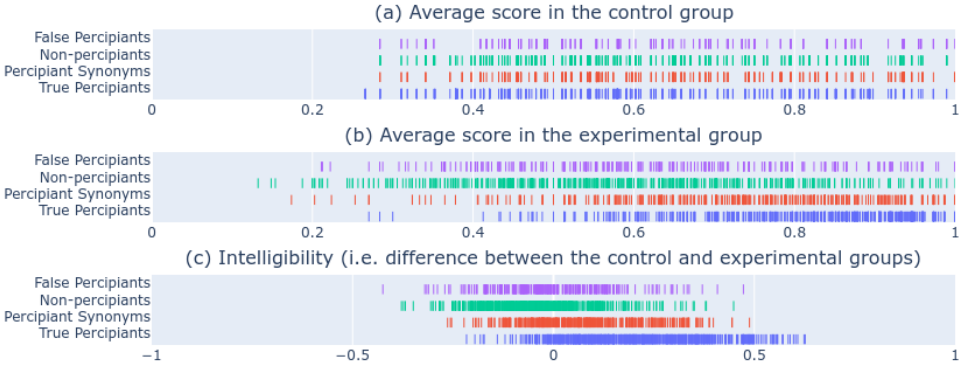


Figure 3: Distribution of the intelligibility of percipiants across the types.

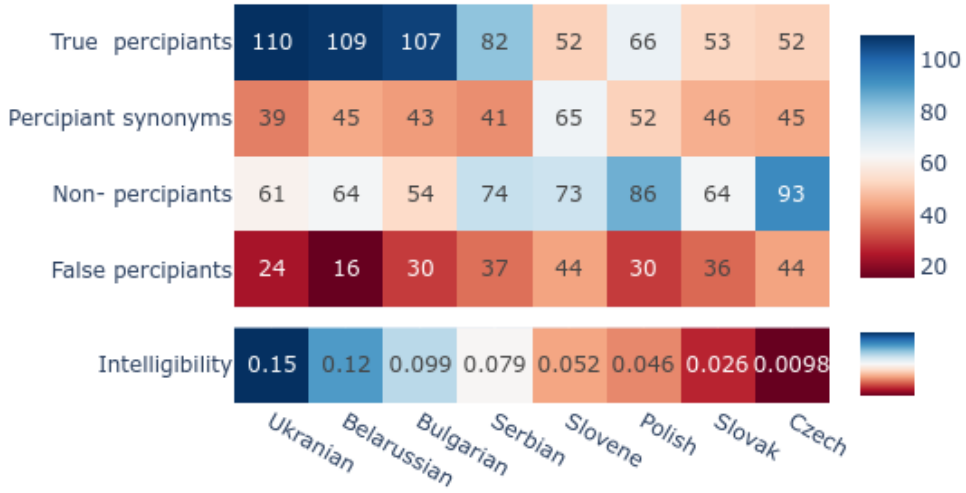


Figure 4: Distribution of the intelligibility of percipiants across the types.

scores (3a and 3b) and intelligibility (3c) across the percipiant classes. The distributions in the control group do not display statistically significant difference (p-values1 between 0.06 and 0.85), except true percipiants vs. percipiant synonyms (p-value<0.001). As for the distribution in the experimental group, it demonstrates a statistically significant difference (p-value<0.001) for each pair of classes. The MI gain relative to the control group is higher for true percipiants. Non-percipiants and false percipiants show the lowest increase, with their mean situated around zero. Percipiant synonyms demonstrate an intermediate behavior. Thus, we conclude that the introduction of the novel MI metric allows differentiating among the classes of percipiants.

Pearson's correlation between MI and the normalized number of true percipiants per language pair (see Figure 4) yielded a strong positive coefficient of 0.9257. We also built a linear regression model to predict MI from the distribution of the four classes, which produced the following coefficients: true percipiants 0.288; percipiant synonyms 0.078; non-percipiants -0.168; false percipiants -0.198; intercept 0.039, R^2 = 0.876. This suggest the following interpretation: true percipiants are the strongest predictors of MI; the predictive power of percipiant synonyms is about four times lower; non-percipiants and false percipiants are about twice as inhibitory as true percipiants are predictive.

5. Conclusion

We introduced the notion of percipiants which are defined as a pair of words which are mutually intelligible within a pair of languages. As distinguished from cognates, percipiants are intentionally agnostic about etymology and common origins. We also defined four classes of percipiants which are differ in the degree of their graphemic and semantic relatedness to their counterparts in genetically related languages. Two experiments were run. The first one analyzed the correctness of predicting missing words under the condition of a parallel text. Similarly to previous work, our analysis attested the general tendency to declining answer correctness with the increase in LD. The second experiment quantified MI as a difference between the average scores from the experimental and the control groups. Such metric enables accurately capturing the net gain in the intelligibility of text with vs. without support from the parallel text. Remarkably, loss in intelligi-

bility is also possible – therefore, parallel text may occasionally become an impediment. The implications of these observations are twofold. Firstly, they reveal a far more pronounced tendency for MI to decrease with higher LD. Secondly, they confirm the validity of dividing percipiants into four classes.

Note, however, that the proposed approach is not free from limitations. On the one hand, the task of restoring an omitted word does not allow isolating the influence exerted by contextual parameters from the perception of individual words. On the other hand, polysemous words may influence the choice of a percipiant pair if stripped of context. MI values with very low or negative gain over the control condition may be explained by genuinely weak MI within the pair, as well as by informational sufficiency of the monolingual context for word prediction. Further experimental and analytical work in this direction is required for a more in-depth understanding of MI and the role of context.

Research ethics: The Institutional Review Board of the National Research University Higher School of Economics at its meeting on August 21, 2023 approved this project as conforming entirely to ethical standards.

References

David Crystal. 2008. *A dictionary of linguistics and phonetics*. Blackwell Publishing.

Jelena Golubović and Charlotte Gooskens. 2015. Mutual intelligibility between West and South Slavic languages. *Russian Linguistics*, 39(3):351–373.

Charlotte Gooskens and Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and change*, 16(3):189–207. Publisher: Cambridge University Press.

Charlotte Gooskens, John Nerbonne, and Nathan Vaillette. 2007. Conditional entropy measures intelligibility among related languages. In *LOT Occasional Series*, volume 7, pages 51–66. Publisher: LOT, Netherlands Graduate School of Linguistics.

Wilbert Jan Heeringa. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University of Groningen, Groningen.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Eduard Klyshinsky. 2021. Statisticheskiĭ metod kolichestvennoĭ otsenki poniatnosti inostrannykh slavianskikh iazykov dlia russkoiazychnogo chitatelia [A statistical method for quantitative evaluation of the intelligibility of foreign Slavic languages for Russian-speaking readers]. *Vestnik Novosibirskogo gosudarstvennogo universiteta*, 19(1):61–79.

Vladimir Levenshtein. 1966. Binary codes capable or correcting deletions, insertions, and reversals. *Soviet Physics–Doklady*, 10(8):707–710.

Koji Miwa, Ton Dijkstra, Patrick Bolger, and R. Harald Baayen. 2014. Reading English with Japanese in mind: Effects of frequency, phonology, and meaning in different-script bilinguals. *Bilingualism: language and cognition*, 17(3):445–463. Publisher: Cambridge University Press.

Tanja C. Roembke, Iring Koch, and Andrea M. Philipp. 2024. What makes a cognate? Implications for research on bilingualism. *Bilingualism: Language and Cognition*, 27(5):1023–1028. Publisher: Cambridge University Press.

Jacopo Saturno. 2023. The impact of interlingual correspondences on cognate recognition in Slavic intercomprehension. *Russian Linguistics*, 47(2):209–229.

Lindsey R. Squires, Sara J. Ohlfest, Kristen E. Santoro, and Jennifer L. Roberts. 2020. Factors influencing cognate performance for young multilingual children's vocabulary: a research synthesis. *American Journal of Speech-Language Pathology*, 29(4):2170–2188.

Irina Stenger, Klára Jágrová, and Tania Avgustinova. 2020. The INCOMSLAV platform: experimental website with integrated methods for measuring linguistic distances and asymmetries in receptive multilingualism. In *Proceedings of the LREC 2020 Workshop on" Citizen Linguistics in Language Resource Development"*, pages 40–48, Marseille, France.

Renée Van Bezooijen and Charlotte Gooskens. 2005. How easy is it for speakers of Dutch to understand Frisian and Afrikaans, and why? *Linguistics in the Netherlands*, 22:13–24.

Guy C. Van Orden. 1987. A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition*, 15(3):181–198.

6. Language Resource References

Klyshinsky and Lukashevich. 2025. *Intelligibility of Slavic Languages for Russian Native Speakers*. Science Data Bank, 1.0.

Transfer Learning for an Endangered Slavic Variety: Dependency Parsing in Pomak Across Contact-Shaped Dialects

Sercan Karakaş
University of Chicago
skarakas@uchicago.edu

Abstract

This paper presents new resources and baselines for Dependency Parsing in Pomak, an endangered Eastern South Slavic language with substantial dialectal variation and no widely adopted standard. We focus on the variety spoken in Turkey (Uzunköprü) and ask how well a dependency parser trained on the existing Pomak Universal Dependencies treebank, which was built primarily from the variety that is spoken in Greece, transfers across dialects. We run two experimental phases. First, we train a parser on the Greek-variety UD data and evaluate zero-shot transfer to Turkish-variety Pomak, quantifying the impact of phonological and morphosyntactic differences. Second, we introduce a new manually annotated Turkish-variety Pomak corpus of 650 sentences and show that, despite its small size, targeted fine-tuning substantially improves accuracy; performance is further boosted by cross-variety transfer learning that combines the two dialects.

Keywords: Pomak, Universal Dependencies, dialectal variation, cross-lingual transfer, low-resource NLP

1. Introduction

Pomak is an endangered Eastern South Slavic language spoken in Bulgaria, Greece, and Turkey, with additional diaspora communities (Constantinides, 2007; Kehaya, 2017; Karakaş, 2020, 2022; Adamou, 2022; Markantonatou et al., 2023). Sociolinguistic syntheses and recent NLP-facing surveys characterize Pomak as severely constrained along multiple vitality dimensions, most prominently intergenerational transmission, functional domain coverage, and institutional/community support dimensions, which closely track UNESCO-style endangerment frameworks (Brenzinger et al., 2003; Markantonatou et al., 2023). From an NLP and resource-construction perspective, Pomak poses a compound low-resource problem: (i) the absence of a widely adopted orthographic standard limits the availability and comparability of written data; and (ii) the speech community forms a dialect continuum whose structural properties have been reshaped by sustained contact with Greek and Turkish in settings of pervasive bi-/trilingualism (Kehaya, 2017; Karakaş, 2022).

More generally, the kinds of lexically and structurally mediated differences observed across Pomak varieties are consistent with well-studied pathways of contact-induced change, in which sustained bilingualism can trigger borrowing, calquing, and convergence effects that extend beyond the lexicon into morphosyntax and argument-structure realizations. Classic work argues that contact outcomes depend on both social factors (intensity, community structure, language attitudes) and linguistic constraints, and that structural transfer can be systematic even in the absence of full language shift (Thomason and Kaufman, 1988; Winford, 2003). In addition, grammatical replication accounts emphasize that contact may promote the diffusion of abstract patterns (e.g., functional-marking strategies) rather than direct morpheme import, yielding typologically coherent but variety-specific reorganizations of form–function mappings (Heine and Kuteva, 2009; Matras, 2009). This perspective motivates treating the Turkey variety not merely as a noisier or less standardized sample of "the same" Pomak, but as a dialect shaped by identifiable contact pressures whose effects are linguistically structured rather than random. On this view, cross-variety differences should not be reduced to superficial lexical substitution or orthographic inconsistency alone, but understood as reflecting contact-conditioned reorganization in the mapping between form and function. This has direct consequences for computational modeling, because it means that cross-variety generalization may fail not simply due to data sparsity, but because the parser is exposed to systematically different distributions of cues in the target variety. It also implies that robust parsing may require a greater amount of in-variety supervision than would be expected if the two varieties differed only minimally. Consequently, variation is not merely lexical but systematically grammatical, affecting phonological inventories, morphosyntactic exponence, and argument-realization options. To give an example, prior descriptive and comparative work documents dialect-sensitive differences in nominal case morphology and oblique marking strategies (Sandry,

[1] Pomak also exhibits dialectal variation within Turkey. This paper is based on the dialect spoken in Uzunköprü, Edirne.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 345–354
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2013), and shows that closely related varieties can diverge in the syntax–morphology mapping for ditransitives (e.g., the encoding of goal/recipient arguments via inflectional case vs. marker/adpositional strategies) (Karakaş, 2022). The contrast in the following sentences exemplifies such micro-variation by comparing the Xanthi (Greece) and Uzunköprü (Turkey) varieties, highlighting consequences for cross-variety generalization in supervised tagging and transfer-based modeling.

(1)

(a) *Xanthi (Greece)*

Ayşe dade resima Aliyu.
Ayşe give.PST.3SG picture-DEF Ali-DAT

'Ayşe gave Ali the picture.'

(b) *Uzunköprü (Turkey)*

Ayşe resim-asa na Ali
Ayşe picture-POSS.DEF DAT.MRK Ali
dade.
give.PST.3SG

'Ayşe gave Ali his/her picture.' [2]

As the example shows, the Xanthi variety licenses an inflectional dative on the recipient (Ali-DAT), whereas the Uzunköprü variety introduces the recipient with *na*, with no overt dative suffix on the proper name. This alternation is consistent with broader claims that Pomak varieties differ in how they encode oblique arguments and distribute case marking across nominal morphology vs. adpositional/marker strategies (Sandry, 2013; Karakaş, 2022).

Given the very limited resources available for Pomak, especially for the variety spoken in Turkey, the study is structured in two phases aimed at developing computational tools to support its processing. The first phase involves building a dependency parser for the Pomak variety spoken in Turkey. This effort builds on the linguistic description provided by Karakaş (2022) for the Turkey variety and on existing morphologically annotated corpora and UD-style resources developed for the Greece variety (Jusúf Karahóğa et al., 2022; Markantonatou et al., 2023). The central research question is whether a parser trained on the Greece variety can effectively accommodate linguistic variation in the Turkey variety, given that Pomak has been documented to show dialectal variation in prior work (Sandry, 2013; Kehaya, 2017; Adamou and Fanciullo, 2018; Adamou, 2022; Karakaş, 2022).

[2]Depending on the variety and analysis, *na* has been treated either as a preposition-like element or as a dedicated marker introducing goal/recipient arguments; we gloss it neutrally as DAT.MRK here.

The second phase of this study introduces a newly developed corpus based on the Pomak variety spoken in Turkey. This corpus serves to evaluate whether the parser can effectively be trained even on a small corpus of the targeted dialect. Due to the limited size of the available data, which results in a poor performance, cross-lingual transfer learning techniques were employed as a solution to overcome the data scarcity. The results demonstrate that cross-lingual transfer learning significantly improves the model's performance, providing the highest accuracy scores when applied to the Pomak dialect which is spoken in Turkey. By addressing these challenges, this study contributes to the growing body of research focused on enhancing computational resources for low-resourced and dialectally diverse languages, with particular emphasis on Pomak.

2. Related Work

Foundational work on Pomak has established both (i) detailed dialect-specific grammatical descriptions and (ii) a growing set of corpus-linked, computationally usable linguistic resources. For the Paševik (Greece) variety, Sandry (2013) provides a broad-coverage description of phonology and nominal/verbal morphology with accompanying notes on core syntax and annotated texts, offering one of the most systematic documentary baselines for Pomak micro-variation. In parallel, generative analyses have shown that dialectal differences are not merely lexical, but can reflect deeper differences in argument structure and syntax–semantics mapping: for example, Karakaş (2022) argues (on the basis of binding, scope, and weak crossover diagnostics) that the Xanthi (Greece) variety patterns with low-applicative double-object structures, whereas the Uzunköprü (Turkey) variety instead aligns with prepositional-ditransitive configurations.

Beyond structural description, recent sociolinguistic work has also clarified how contact ecologies and speaker attitudes bear directly on maintenance. In particular, Dağdeviren-Kırmızı (2026) investigates linguistic (in)security among Pomak heritage speakers in Western Türkiye via semi-structured interviews, reporting systematic age- and language-contingent proficiency profiles and pervasive insecurity across interactional settings. Importantly for resource-building, the study links family language practices and failures of intergenerational transmission to experienced (in)security and to subtractive schooling effects in the dominant language, which may show that data scarcity is partly socially produced rather than purely technological.

In parallel, there is rapidly growing NLP-oriented interest in Pomak, with recent work targeting the creation of reusable digital infrastructures and open

resources Krimpas et al. (2021). On the text side, Jusúf Karahóǧa et al. (2022) describe the development of a Latin-based orthography and gold morphological annotation for Pomak, explicitly aligning the resulting resources with UD-style dissemination and downstream tool training. Complementing this, Kokkas et al. (2025) report the digitization and exploration of a Pomak corpus compiled in the *Philotis* project (2020–2023), comprising roughly 140k words across diverse genres and released as an open textual resource intended for NLP practitioners and quantitative corpus analysis. Extending beyond text, Tsoukala et al. (2023) present an end-to-end pipeline for training ASR systems for low-resourced languages and use Pomak as a case study, training what they describe as the first Pomak ASR model. These efforts motivate a modeling agenda in which cross-variety transfer, orthography/annotation decisions, and modality-specific pipelines are treated as first-class design variables for Pomak NLP.

3. Methods

To assess whether Pomak's documented dialectal heterogeneity translates into measurable degradation in syntactic generalization, we fine-tuned a state-of-the-art graph-based neural dependency parser (Dozat et al., 2017) on the Pomak Universal Dependencies (UD) treebank (Jusúf Karahóǧa et al., 2022). The treebank is derived from the QPM corpus with syntactic annotation guided by an active-learning workflow (Settles, 2009; Anastasopoulos et al., 2018; Shi et al., 2021). Concretely, the active annotation strategy iteratively prioritizes informative sentences for manual correction, thereby concentrating scarce expert effort on examples expected to yield the largest model and guideline improvements, which can be considered to be a particularly appropriate design choice for a non-standardized, low-resource language with substantial intra-variety variation.

All parsing experiments use our own implementation of the biaffine, graph-based dependency parser of Dozat et al. (2017). The reason why we adopt a graph-based parser is because it directly optimizes over full dependency graphs and has been shown to be robust for morphologically rich languages, where fusional exponence and flexible word order increase the density of long-distance dependencies and raise ambiguity for locally greedy transition systems (Dozat et al., 2017). This architectural bias is especially relevant for Pomak, whose Slavic-type morphology and contact-shaped micro-variation can interact to produce non-trivial argument-structure and case-marking patterns.

3.1. Parser Architecture

We implement the biaffine, graph-based dependency parser of Dozat et al. (2017). Given a sentence of tokens $w_{1:n}$, the model (i) builds contextual token representations, (ii) scores all possible head–dependent arcs with a biaffine scorer, (iii) scores dependency relation labels for each chosen arc, and (iv) decodes the highest-scoring well-formed dependency tree (non-projective maximum spanning arborescence) and assigns labels.

Token representations and contextual encoder. Each token w_i is represented by a concatenation of a word embedding and a character-composed embedding (optionally augmented by discrete features):

(2) Input representation.

$$x_i = \left[e_i^{(w)};\ e_i^{(c)};\ e_i^{(\text{feat})}\right],$$
$$e_i^{(c)} = \text{CharEnc}(\text{ch}_{i,1:m_i}).$$

Contextualization.

$$r_i = \text{BiLSTM}(x_{1:n})_i.$$

Arc scoring (biaffine). We produce separate head and dependent projections for arc prediction using MLPs, then score every candidate head j for dependent i:

(3) Arc-specific projections.

$$d_i^a = \text{MLP}^{\text{dep},a}(r_i), \qquad h_j^a = \text{MLP}^{\text{head},a}(r_j).$$

Biaffine arc score.

$$s_{ij}^a = (d_i^a)^\top U^a h_j^a + (u^a)^\top \left[d_i^a; h_j^a\right] + b^a.$$

Head distribution.

$$p(h_i = j \mid i) = \text{softmax}_j\left(s_{ij}^a\right),$$

where h_i denotes the head index of token i (including a designated ROOT).

Relation labeling (biaffine). Conditioned on a head choice j for dependent i, we score a distribution over relation labels $\ell \in \mathcal{R}$:

(4) Relation-specific projections.

$$d_i^r = \text{MLP}^{\text{dep},r}(r_i), \qquad h_j^r = \text{MLP}^{\text{head},r}(r_j).$$

Biaffine relation score (vector over labels).

$$s_{ij}^r = (d_i^r)^\top U^r h_j^r + W^r\left[d_i^r; h_j^r\right] + b^r,$$

where $s_{ij}^r \in R^{|\mathcal{R}|}$, U^r is a label-tensor, and W^r, b^r are label-wise affine parameters.

Relation distribution.

$$p(r_i = \ell \mid i, h_i = j) = \text{softmax}_\ell\left(s_{ij}^r[\ell]\right),$$

where r_i denotes the relation label of token i.

Decoding. At inference time, we decode a single well-formed dependency tree using the arc scores $\{s^a_{ij}\}$ (allowing non-projective structures) and then assign a label to each selected arc:

(5) Tree decoding (arcs).

$$\hat{T} = \arg\max_{T \in \mathcal{T}} \sum_{(j \to i) \in T} s^a_{ij},$$

where $\mathcal{T}$ is the set of valid dependency trees (arborescences). Label assignment.

$$\hat{r}_i = \arg\max_{\ell \in \mathcal{R}} s^r_{i\,\hat{h}_i}[\ell],$$

with $\hat{h}_i$ given by the decoded tree $\hat{T}$.

Training objective. Let $j^\star_i$ be the gold head of token i and $\ell^\star_i$ the gold relation label. We minimize the negative log-likelihood of (i) head selection and (ii) relation labeling:

(6) Arc loss.

$$\mathcal{L}_{\text{arc}} = -\sum_{i=1}^{n} \log p(h_i = j^\star_i \mid i)$$

Relation loss.

$$\mathcal{L}_{\text{rel}} = -\sum_{i=1}^{n} \log p(r_i = \ell^\star_i \mid i, h_i = j^\star_i)$$

Total objective.

$$\mathcal{L} = \mathcal{L}_{\text{arc}} + \mathcal{L}_{\text{rel}} + \lambda \|\theta\|_2^2.$$

This formulation cleanly separates head selection (UAS-relevant) from relation labeling (LAS-relevant), matching the evaluation in Section 4.

3.2. A New Corpus

Recall that the second phase of the study addresses a central limitation of existing Pomak NLP resources: the lack of dialect-matched annotated data for the variety spoken in Turkey. Although the currently available Pomak UD treebank has made initial modeling possible, it mainly reflects the Greece variety and thus can be said to capture lexical choices, morphological patterns, and syntactic distributions that are not always representative of the Uzunköprü variety. This matters because dialectal differences are not limited to pronunciation or a small number of vocabulary items as we established in the previous sections. They can also affect the forms of case marking, the realization of verbal morphology, the choice between competing function words or adpositions, and the distribution of dependency relations in context. In the Uzunköprü variety, these contrasts are further shaped by sustained contact with Turkish, which may influence both lexical selection and morphosyntactic preferences. As a result, a model trained only on the Greece-variety treebank may learn analyses that are well suited to that variety but systematically mis-handle tokens and structures that are natural in Uzunköprü. Creating dialect-matched annotated data is therefore necessary not only for improving accuracy, but also for testing how well current NLP methods transfer across closely related yet contact-shaped Pomak varieties. This is particularly consequential in Pomak, where sustained bilingualism and long-term contact with Turkish can yield contact-induced divergence in both the lexicon and the grammar, affecting the surface realization of functional material and argument-marking strategies (Karakaş, 2022; Thomason and Kaufman, 1988; Winford, 2003; Heine and Kuteva, 2009; Matras, 2009). In low-resource settings, such distributional shifts are amplified: even small differences in case/marker realizations, cliticization patterns, or function-word inventories can systematically alter the feature space that a tagger or parser relies on, thereby degrading cross-variety transfer.

To enable controlled evaluation and adaptation, we compiled a mini corpus of the Pomak variety spoken in Turkey, consisting of 650 sentences drawn from two complementary sources. First, we collected sentence-level data from six native speakers of the Turkey variety of Pomak living in Turkey. These speakers were between 47 and 60 years old and were bilingual in Pomak and Turkish. The speaker-contributed portion of the corpus primarily reflects everyday language use rather than highly formal, literary, or specialized registers. Second, we incorporated additional sentences from a study that provides primary data on the Turkey variety (Karakaş, 2022). Therefore, the corpus should be understood as a small mixed-source sentence-level resource rather than a genre-balanced corpus.

Because Pomak lacks a widely adopted orthographic standard, corpus preparation involved an explicit normalization and annotation workflow. We first normalized orthographic forms and standardized tokenization so that the data could be used consistently for supervised modeling and compared directly with existing UD-style Pomak resources. We then annotated the corpus in accordance with the Universal Dependencies framework, assigning POS tags, morphological features, head–dependent relations, and dependency labels in a UD-compatible format. This workflow was designed to ensure maximal comparability between the existing Greece-variety UD resource and the newly compiled Turkey-variety corpus, while also making the new dataset suitable for transfer-based parsing experiments.

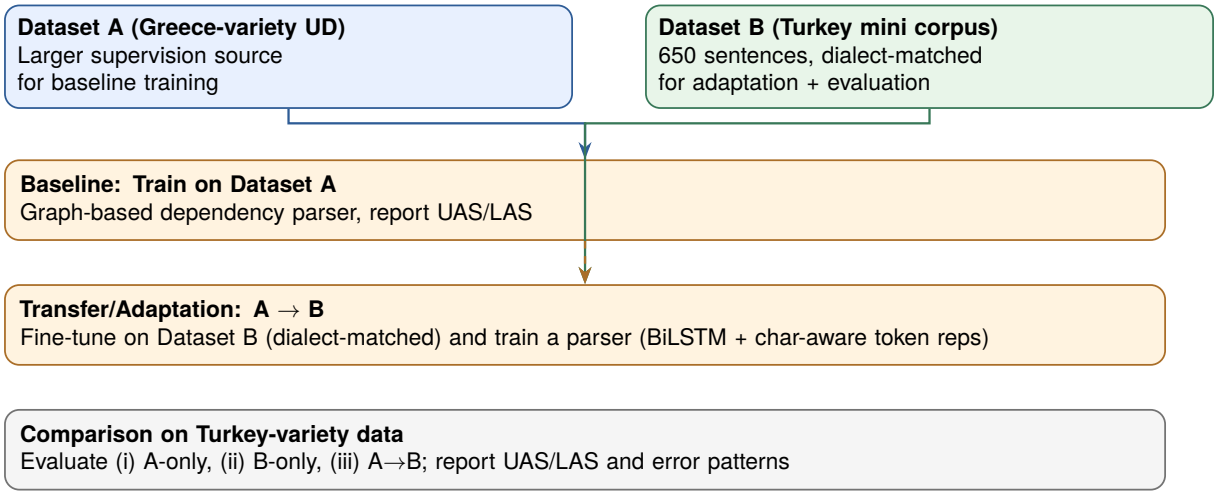


Figure 1: Workflow summary. Dataset A provides baseline supervision; Dataset B provides dialect-matched data for adaptation and evaluation.

Despite its usefulness, the corpus size is intentionally small and therefore represents the "extreme low-resource" regime that characterizes endangered-language NLP in practice. When we train the syntactic pipeline exclusively on this 650-sentence corpus, performance is substantially lower than models trained on the larger Greece-variety UD resource, reaching only 50 UAS and 44 LAS. This result is informative in two respects. First, it quantifies the limits of purely within-variety supervision at this scale: even with dialect-matched data, the model cannot learn sufficiently robust generalizations from 650 sentences. Second, it motivates the transfer-learning strategy pursued in later sections, where cross-variety supervision from the larger Pomak UD resource is combined with targeted fine-tuning on the Turkey-variety mini corpus. In other words, the mini corpus is not intended to replace existing resources, but to serve as a crucial *adaptation set* that anchors the model in the distributional and contact-shaped properties of the Turkey variety while leveraging larger out-of-variety supervision for statistical strength.

We distinguish three datasets. Dataset A is the existing Pomak UD treebank (primarily Greece variety). Dataset B is our newly compiled Turkey-variety corpus of 650 sentences, drawn from bilingual Pomak–Turkish speakers living in Turkey and from prior descriptive materials, which we split 80/20 into 520 sentences for training and 130 sentences for development (early stopping and model selection). Finally, Dataset C is a held-out Turkey-variety test set of 90 sentences collected from a *different native speaker* than those contributing to Dataset B (speaker-disjoint evaluation). All reported Turkey-variety results (Tables 1–3) are computed on Dataset C, ensuring comparability across training regimes.

4. Evaluation Metrics and Results

We evaluate syntactic parsing performance using two standard dependency-parsing metrics: unlabeled attachment score (UAS) and labeled attachment score (LAS). UAS measures the proportion of tokens whose predicted head matches the gold head, abstracting away from relation type. LAS is stricter: a token is counted as correct only if *both* the predicted head and the dependency relation label match the gold annotation. Throughout, scores are computed at the word level and exclude punctuation following common UD evaluation conventions.

Table 1 reports baseline cross-variety performance when the parser is trained on the Greece-variety Pomak UD resource and evaluated on Turkey-variety Pomak. The baseline achieves 59% UAS and 51% LAS. Two aspects of this pattern are noteworthy. First, the gap between UAS and LAS (8 points) indicates that a non-trivial share of errors involve relation labeling *in addition to* head selection, consistent with the possibility that dialect-conditioned differences affect not only local attachment preferences but also the distribution of functional material and morphosyntactic cues used to disambiguate relation types. Second, the overall level of performance is substantially below what is typically observed when training and testing are conducted within the same variety under comparable annotation conditions, suggesting that cross-variety transfer is far from automatic even within a closely related continuum.

Interpreting these results in light of Pomak's documented micro-variation, the baseline scores are consistent with a distribution-shift account: the Greece-variety training data may under-represent (or encode differently) constructions that are frequent in the Turkey variety, including contact-

Setup	Metric	Score
Train: Dataset A; Test: Dataset C	UAS	59%
Train: Dataset A; Test: Dataset C	LAS	51%

Table 1: Baseline cross-variety parsing performance. Dataset A is the existing Pomak UD treebank (primarily Greece variety), and Dataset C is the held-out Turkey-variety test set from a speaker not represented in Dataset B.

shaped argument-marking strategies, function-word inventories, and morphosyntactic exponence patterns. In low-resource settings, such shifts are amplified because the model has limited opportunity to learn variety-invariant abstractions, and instead relies heavily on surface cues whose realizations differ across dialects. The baseline therefore provides direct quantitative evidence that the existing UD resource does not straightforwardly "cover" the Turkey variety, motivating the dialect-matched mini corpus and the adaptation experiments presented in the following section.

4.1. Turkey-Variety Mini Corpus and Transfer Results

Recall that in order to assess how much *dialect-matched* supervision can mitigate cross-variety degradation, and whether transfer-based training can further improve generalization under extreme data scarcity, we compare three training regimes. The first regime is the cross-variety baseline: the model is trained on the existing Pomak UD resource, which primarily reflects the Greece variety, and evaluated on Turkey-variety Pomak. The second regime is an in-variety condition in which the model is trained exclusively on our newly compiled Turkey-variety mini corpus. The third regime introduces transfer/adaptation: we leverage the broader syntactic coverage available in the larger UD dataset while anchoring the model in the Turkey-variety distribution via fine-tuning on the mini corpus.

Table 2 compares the cross-variety baseline against the in-variety (mini-corpus-only) condition. Despite being dialect-matched, the mini-corpus-only model attains 50% UAS and 44% LAS, underperforming the model trained on the larger Greece-variety UD resource (59% UAS / 51% LAS). This outcome highlights the severity of the data-scarcity regime: 650 sentences do not provide sufficient coverage for the model to learn robust syntactic generalizations, particularly for lower-frequency constructions and long-distance dependencies. Importantly, this result separates two distinct constraints. Dialect match reduces distribution shift, but it cannot compensate for the statistical advantages of substantially larger supervision. In other words, within-variety training at this scale remains bottlenecked by coverage rather than by dialect mismatch alone.

Training / Evaluation Setup	UAS	LAS
Dataset A → Dataset C	59%	51%
Dataset B-train → Dataset C	50%	44%

Table 2: Cross-variety baseline versus in-variety training. Dataset A is the existing Pomak UD treebank (primarily Greece variety). Dataset B-train and B-dev are splits of our Turkey-variety corpus, and Dataset C is the held-out Turkey-variety test set from a different speaker.

We therefore evaluate a transfer/adaptation regime that combines the strengths of both resources: the larger UD dataset supplies broader syntactic coverage and reduces estimation variance, while the mini corpus provides dialect-specific calibration for the target variety. Intuitively, the transfer setup encourages the model to learn higher-level abstractions from the larger resource and then adjust the mapping from surface cues to syntactic structure in the Turkey variety during fine-tuning. Under this regime, performance improves to 68% UAS and 62% LAS on a held-out Turkey-variety test set of 90 sentences (Table 3). The size of this gain relative to both baselines suggests that additional supervision can be exploited without sacrificing dialect specificity, provided that training explicitly incorporates an adaptation step. Notably, the improvement is observed in both UAS and LAS, indicating that adaptation helps not only with head selection but also with relation labeling—components that are plausibly sensitive to contact-shaped differences in functional marking and argument encoding.

Overall, the results support a practical methodological conclusion for endangered languages with strong dialectal structure: when dialect-matched corpora are necessarily small, the most effective strategy is often not to train from scratch on the target dialect, but to treat dialect-matched data as an *adaptation resource* integrated into a broader transfer pipeline. In our case, the transfer regime substantially improves parsing quality on the Turkey variety, suggesting a path forward for building usable syntactic tools even when in-variety annotation is limited.

5. Discussion

Our results provide converging evidence that Pomak NLP must be treated as a *dialect-sensitive* low-resource problem rather than a single homogeneous target. Training on the existing Pomak

Model / Training Regime	UAS	LAS
Original Pomak UD (Greece variety)	59%	51%
Mini corpus	50%	44%
Transfer / adaptation (A→B)	68%	62%

Table 3: Parsing performance under three regimes: out-of-variety baseline, in-variety mini-corpus-only training, and transfer/adaptation leveraging both resources.

UD resource (primarily Greece variety) yields only moderate cross-variety generalization to the Turkey variety (59 UAS / 51 LAS), indicating a substantial distributional mismatch despite genealogical relatedness. This pattern is consistent with the linguistic literature portraying Pomak as a contact-shaped dialect continuum: sustained bilingualism and long-term contact with Turkish and Greek can induce both lexical turnover and systematic morphosyntactic reorganization, including shifts in functional inventories and argument-encoding strategies as also shown in (Karakaş, 2022). In such circumstances, models trained on one variety may internalize variety-specific correlates of syntactic relations—for example, the co-occurrence patterns between function words, case/marker realizations, and dependency labels—that do not transfer faithfully to another variety, yielding errors in both attachment (UAS) and relation assignment (LAS). The non-trivial UAS–LAS gap in the baseline further suggests that relation labeling is particularly sensitive to these shifts, plausibly because UD relation inventories are mapped from surface cues that may change under contact (e.g., oblique marking strategies, adpositional elements, or the distribution of possessive morphology).

At the same time, the comparison between in-variety and transfer-based training shows that *dialect match alone is not sufficient* when labeled supervision is extremely limited, echoing core results in domain and parser adaptation that low-resource target settings typically benefit from transfer and semi-/self-supervised adaptation (Daumé III, 2007; Jiang and Zhai, 2007; McClosky et al., 2006; Gururangan et al., 2020). Training exclusively on the 650-sentence Turkey-variety mini corpus leads to lower performance (50 UAS / 44 LAS) than the cross-variety baseline, despite being distributionally better aligned with the test variety. This outcome may suggest a central tension for endangered-language NLP, which is that the very scenarios that most demand dialect-matched resources are those where annotation budgets are smallest, and small in-variety corpora may lack the coverage required to estimate robust syntactic generalizations. In practice, the model must generalize across sparse constructions, long-distance dependencies, and low-frequency lexical types; with only a few hundred sentences, learning is dominated by estimation variance and incomplete coverage rather than by mismatch alone. This result also suggests that, for Pomak specifically, dialect-conditioned lexical/functional variability may interact with orthographic non-standardization to inflate the effective type inventory, further increasing sparsity under small corpora.

Crucially, the transfer/adaptation regime resolves this tension by combining broad-coverage supervision from the larger UD resource with dialect-specific calibration from the mini corpus. Fine-tuning/adaptation improves performance to 68 UAS / 62 LAS on the Turkey-variety test set, yielding gains over both baselines. We interpret this improvement as evidence that (i) the larger resource supplies essential syntactic coverage and stabilizes parameter estimates, while (ii) the dialect-matched mini corpus effectively "reanchors" the learned representations in the target variety, adjusting the mapping from surface diagnostics to syntactic structure. That the improvement appears in both UAS and LAS is particularly important: adaptation helps not only with head selection but also with relation labeling, implying that the fine-tuning signal corrects systematic mismatches in how functional material and argument marking correlate with UD relations in the Turkey variety. Methodologically, these findings align with prior work showing that multi-treebank/multilingual supervision provides broad syntactic coverage for low-resource parsing and that an additional fine-tuning/adaptation step on limited target-variety data can substantially improve both unlabeled and labeled attachment, effectively calibrating the parser to the target distribution (Stymne et al., 2018; Kondratyuk and Straka, 2019; Stymne, 2020; Glavaš and Vulić, 2021; Velldal et al., 2017; Abo Mokh et al., 2024).

The results also have implications for resource design. First, they argue for explicitly encoding *variety metadata* in Pomak resources (e.g., dialect tags, speaker region, orthographic conventions), since the performance gap demonstrates that pooling across varieties without control can obscure distribution shifts that matter for modeling. Second, they motivate targeted annotation strategies that prioritize constructions likely to differ across varieties—for example, oblique marking and ditransitives, functional elements like *na*-type markers, and contact-sensitive possessive and case morphology. In future work, this could be operationalized via active learning criteria that are explicitly variety-aware (e.g., uncertainty sampling over dialect-diagnostic contexts), thereby increasing the return on limited annotation effort. Finally, while our experiments focus on dependency parsing, the

same logic extends to downstream tasks (information extraction, morphology, ASR/MT): contact-conditioned dialect differences can create systematic error modes that are not reducible to "noise" and therefore benefit from explicit adaptation.

Several limitations point to concrete next steps. The Turkey-variety corpus remains small, and our evaluation set should be expanded to reduce uncertainty and enable more fine-grained error analyses by construction type. In addition, orthographic and tokenization choices are likely to interact with character-level modeling; systematically comparing normalization schemes and subword segmentation strategies may further improve robustness. Finally, more diagnostic analyses—e.g., confusion matrices over POS and dependency labels, and targeted evaluation on constructions known to differentiate varieties—would sharpen the linguistic interpretation of the observed gains and help identify which aspects of contact-shaped variation are most consequential for syntactic modeling.

6. Conclusion

This paper investigated whether Pomak's contact-shaped dialectal variation poses a measurable challenge for syntactic processing and whether transfer-based modeling can mitigate this challenge under severe data scarcity. Using the existing Pomak UD resource (primarily representing the Greece variety), we found only moderate cross-variety generalization to the Turkey variety, consistent with a substantial distribution shift in the surface cues that support both attachment and relation labeling. To directly probe this mismatch, we constructed a Turkey-variety mini corpus of 650 sentences compiled from six native speakers and curated thesis data, yielding UD-compatible supervision explicitly targeting the Uzunköprü variety.

Our experiments show that dialect match alone is not sufficient at this scale: training exclusively on the mini corpus underperforms training on the larger out-of-variety UD resource, highlighting the limits of purely in-variety supervision in the extreme low-resource regime. By contrast, a transfer/adaptation strategy that combines broad-coverage supervision from the larger UD resource with fine-tuning on the dialect-matched mini corpus yields the strongest results, improving both UAS and LAS on Turkey-variety evaluation. Taken together, these findings support a practical methodological recommendation for endangered, dialectally diverse languages: when dialect-specific corpora are necessarily small, they are most effective when used as an adaptation signal integrated into a broader transfer pipeline, rather than as a stand-alone training resource.

7. Limitations and Future Directions

Several limitations of the present study point to clear directions for future work. First, the current analysis does not yet isolate parser behavior on dialect-sensitive constructions that are especially relevant for Pomak, such as oblique-marking strategies, ditransitives, case-marking asymmetries, or function-word realizations shaped by contact. As a result, while the quantitative comparison demonstrates that cross-variety transfer is imperfect and that fine-tuning improves performance, the present paper cannot yet determine which specific grammatical phenomena are driving the observed degradation and subsequent gains. Besides, the Turkey-variety corpus remains small, and this necessarily constrains both modeling and evaluation. Although the 650-sentence corpus is sufficient to demonstrate the value of dialect-matched adaptation, it is too limited to support fine-grained analysis across a broad range of constructions, genres, and speaker profiles. Future work should therefore expand the Turkey-variety dataset substantially, both by increasing the number of annotated sentences and by broadening the range of speakers and usage contexts represented in the corpus.

8. Acknowledgements

I am deeply grateful to Balkız Öztürk for her guidance on endangered languages. I also thank Emirhan Özkan for his help in reaching the Pomak language community. Finally, I sincerely thank the Pomak speakers who generously shared their language with me; this work would not have been possible without their time, trust, and knowledge.

9. Bibliographical References

N. Abo Mokh, D. Dakota, and S. Kübler. 2024. Out-of-domain dependency parsing for dialects of Arabic: A case study. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 170–182, Bangkok, Thailand. Association for Computational Linguistics.

E. Adamou. 2022. Pomak. Encyclopedia entry in *Encyclopedia of Slavic Languages and Linguistics (ESLL)*. Author manuscript / record available via HAL.

E. Adamou and D. Fanciullo. 2018. Why Pomak will not be the next Slavic literary language. In *Linguistic Regionalism in Eastern Europe and Beyond: Minority, Regional and Literary Microlanguages*, pages 40–65. Available online via HAL.

A. Anastasopoulos, M. Lekakou, J. Quer, E. Zimianiti, J. DeBenedetto, and D. Chiang. 2018. Part-of-speech tagging on an endangered language: A parallel Griko–Italian resource. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2529–2539.

M. Brenzinger, A. Yamamoto, N. Aikawa, D. Koundiouba, A. Minasyan, A. Dwyer, C. Grinevald, M. Krauss, O. Miyaoka, O. Sakiyama, et al. 2003. Language vitality and endangerment. UNESCO document, Safeguarding Endangered Languages.

N. T. Constantinides. 2007. Units of the Pomak civilization in Greek Thrace: Brief historical review, language and identities. Master's thesis, Democritus University of Thrace.

G. Dağdeviren-Kırmızı. 2026. Linguistic (in)security perceptions and family language policy among pomak speakers in western türkiye. *Language Policy*, 25(6). Article number 6; published 28 January 2026.

H. Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

T. Dozat, P. Qi, and C. D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

G. Glavaš and I. Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online. Association for Computational Linguistics.

S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

B. Heine and T. Kuteva. 2009. *Language Contact and Grammatical Change*. Cambridge University Press, Cambridge.

J. Jiang and C. Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic. Association for Computational Linguistics.

R. Jusúf Karahóğa, P. G. Krimpas, V. Stamou, V. Arampatzakis, D. Karamatskos, V. Sevetlidis, N. Constantinides, N. Kokkas, G. Pavlidis, and S. Markantonatou. 2022. Morphologically annotated corpora of Pomak. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 179–186, Dublin, Ireland. Association for Computational Linguistics.

S. Karakaş. 2020. Low and high applicatives in Pomak. Abstract, Heritage Language Syntax (HLS1) virtual conference (Utrecht University / ERC Microcontact), 12–16 Oct 2020.

S. Karakaş. 2022. Applicatives in Pomak. Master's thesis, Boğaziçi University.

S. A. Kehaya. 2017. An analysis of the Pomak language based on fieldwork research data. Master's thesis, Aristotle University of Thessaloniki. Available online (institutional repository PDF).

N. T. Kokkas, S. Markantonatou, and V. Arampatzakis. 2025. Exploration and digitization of the pomak linguistic corpus. In G. Pavlidis and S. Sylaiou, editors, *Transforming Heritage Research in a Transforming World: 5th CAA-GR Conference 2024*, Springer Proceedings in Archaeology and Heritage, pages 407–415. Springer, Cham. First Online: 02 January 2026.

D. Kondratyuk and M. Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

P. G. Krimpas, N. Konstantinidis, R. Karahoğa, S. Markantonatou, and G. Pavlidis. 2021. Pomak: An idiosyncratic East South Slavic Language? In *7th International Scientific Conference on Greek Traditional Culture: Conference Proceedings*, pages 66–70, Moscow. Izdatel'stvo Moskovskogo universiteta.

S. Markantonatou, N. T. Constantinides, V. Stamou, V. Arampatzakis, P. G. Krimpas, and G. Pavlidis. 2023. Methodological issues regarding the semi-automatic UD treebank creation of under-resourced languages: The case of Pomak. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 27–35.

Y. Matras. 2009. *Language Contact*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

D. McClosky, E. Charniak, and M. Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 337–344, Sydney, Australia. Association for Computational Linguistics.

S. Sandry. 2013. *Phonology and Morphology of Paševik Pomak with Notes on the Verb and Fundamentals of Syntax*. Phd thesis, University College London. Available via UCL Discovery repository.

B. Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. Updated on January 26, 2010.

T. Shi, A. Benton, I. Malioutov, and O. Irsoy. 2021. Diversity-aware batch active learning for dependency parsing.

S. Stymne. 2020. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany. Association for Computational Linguistics.

S. Stymne, M. de Lhoneux, A. Smith, and J. Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.

S. G. Thomason and T. Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley.

C. Tsoukala, K. Kritsis, I. Douros, A. Katsamanis, N. Kokkas, V. Arampatzakis, V. Sevetlidis, S. Markantonatou, and G. Pavlidis. 2023. ASR pipeline for low-resourced languages: A case study on pomak. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 40–45, Dubrovnik, Croatia. Association for Computational Linguistics.

E. Velldal, L. Øvrelid, and P. Hohle. 2017. Joint UD parsing of Norwegian Bokmål and Nynorsk. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

D. Winford. 2003. *An Introduction to Contact Linguistics*. Wiley-Blackwell, Oxford.

10. Language Resource References

R. Jusúf Karahóğa, P. G. Krimpas, V. Stamou, V. Arampatzakis, D. Karamatskos, V. Sevetlidis, N. Constantinides, N. Kokkas, G. Pavlidis, and S. Markantonatou. 2022. Morphologically annotated corpora of Pomak. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 179–186, Dublin, Ireland. Association for Computational Linguistics.

S. Markantonatou, N. T. Constantinides, V. Stamou, V. Arampatzakis, P. G. Krimpas, and G. Pavlidis. 2023. Methodological issues regarding the semi-automatic UD treebank creation of under-resourced languages: The case of Pomak. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 27–35.

Author Index

Abdulazeez, Dezheen H., 110
Afanasev, Ilia, 195
Agerri, Rodrigo, 153
Ahmadi, Sina, 135
Akhter, Maria Hassan, 83
Alabi, Jesujoba, 12
Alam, Md. Golam Rabiul, 238
Anagnostopoulou, Elena, 287
Anastasopoulos, Antonios, 308
Armostis, Spyros, 308
Artioli, Andrea, 70

Badryzlova, Yulia, 337
Bengoetxea, Jaione, 153
Berikashvili, Svetlana, 230
Bigeard, Sam, 256
Blevins, Margaret, 58
Blevins, Margaret M., 221
Boas, Hans C., 221
Bompolas, Stavros, 287, 315
Bystrich, Tobias, 83

Calderaro, Silvio, 1
Carletti, Lorenzo, 70
Chatzikyriakidis, Stergios, 265
Chousou-Polydouri, Natalia, 287
Coats, Steven, 24
Cutura, Rene, 165

Dheyab, Abuobaida Abdullah, 110
Doyle, Adrian, 33
Ducceschi, Luca, 188

E Sobhani, Mahbub, 238
Erhart, Pascale, 256

Ferroni, Sofia, 70
Fischbach, Lea, 83, 93, 177
Fischer, Hanna, 144
Flek, Lucie, 83, 93, 177
Franzini, Greta H., 188
Frund, Yuliia, 135

Gagnier, Henry, 124
Genitsaridi, Manuela, 287
Gilbert, Glenn, 221
Giouli, Voula, 300
Gonzalez-Dios, Itziar, 153
Guardiano, Cristina, 70

Hamm, Lucile, 256
Hamm, Lukas, 83
Hassan, Md. Rezuwan, 238
Henkel, Maximilian, 165
Hernaez, Inma, 48
Herranz, Asier, 48

Jaff, Daban Q., 104

Karakas, Sercan, 345
Karantzola, Eleni, 265
Karatzanou, Danae, 287
Karimi, Akbar, 83, 177
Katsouros, Vassilis, 308
Klakow, Dietrich, 12
Kleen, Caroline, 177
Klimi, Antigoni, 308
Klyshinsky, Edward, 337
Konstantinidou, Rodanna, 300
Kostopoulos, Georgios, 287

Lameli, Alfred, 93, 144, 177
Lekakou, Marika, 315
Li Destri, Emanuela, 70
Longhin, Marco, 70
Longobardi, Giuseppe, 70

Makri, Vasiliki, 265
Markantonatou, Stella, 300, 315
Marongiu, Paola, 274
Matteazzi, Giovanni Battista, 70
Melhum, Amera Ismail, 110
Melissaropoulou, Dimitra, 287
Miletić, Filip, 165
Mohammadamini, Mohammad, 104, 110
Mohammed, Aveen Jalal, 110
Mohammed, Barzan Hussein, 110
Montemagni, Simonetta, 274
Monti, Johanna, 1
Motta, Federico, 70

Nastevska, Hristina, 327
Navas, Eva, 48
Ngo, Quynh Quang, 165
Niloy, Soumik Deb, 238

Oberkircher, Lena Sophie, 12
Ouni, Slim, 256

Paraskevopoulos, Georgios, 308
Pendevska, Marija, 327
Pinna, Emanuela, 315

Quer, Josep, 315

Rahman, Subhey Sadi, 238

Sadeeq, Imad Saeed, 110
Sadeque, Farig Yousuf, 238
Sadler, Sophie, 165
Salih, Dilgash Mohammed, 110
Saratxaga, Ibon, 48
Schmidt, Thomas, 221
Schulte im Walde, Sabine, 165
Sedlmair, Michael, 165
Skopeteas, Stavros, 230
Sorge, Gaia, 70
Souganidis, Christoforos, 48
Stamou, Vivian, 300, 308

Trouvain, Jürgen, 12
Tsoukala, Chara, 300

van Dam, Kellen Parker, 116

Werner, Carole, 256

Yaich, Malek, 256

Zaghouani, Wajdi, 209