LREC 2026

The Sixth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID@MENTAL.ai) @ LREC 2026

Workshop Proceedings

Editors
Dimitrios Kokkinakis, Charalambos Themistocleous, Gaël Dias, Kathleen C. Fraser, Sebastião Pais, Fredrik Öhman

12 May 2026

Proceedings of Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments - In collaboration with the MENTAL.ai project (RaPID-6@MENTAL.ai 2026) @ LREC 2026

ISBN 978-2-493814-59-3

Message from the General Chair

Welcome to the LREC 2026 Workshop on *"Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments"*, this year in a cooperation with the *"MENTAL.ai consortium"* (RaPID-6@MENTAL.ai). This volume documents the Proceedings of the RaPID-6@MENTAL.ai Workshop, held on Tuesday afternoon, May 12th, 2026, as part of the 15th edition of the LREC 2026 conference *"the International Conference on Language Resources and Evaluation"*.

RaPID-6@MENTAL.ai aims to be an interdisciplinary forum for researchers to share information, findings, methods, models and experience on the collection and processing of data produced by people with various forms of mental, cognitive, neuropsychiatric, or neurodegenerative impairments, such as aphasia, dementia, autism, bipolar disorder, Parkinson's disease, or schizophrenia. Like the previous five editions, the RaPID-6@MENTAL.ai workshop's focus is on creation, processing, and application of data resources from individuals at various stages of these impairments and with varying degrees of severity. Creation of resources includes e.g. annotation, description, analysis, and interpretation of linguistic, paralinguistic and extra-linguistic data (such as spontaneous spoken language, transcripts, eye tracking measurements, wearable and sensor data, etc). Processing is done to identify, extract, correlate, evaluate and disseminate various linguistic or multimodal phenotypes and measurements, which then can be applied to aid diagnosis, monitor the progression, or predict individuals at risk.

RaPID-6@MENTAL.ai invited submissions of papers in all of the aforementioned research areas, particularly emphasizing the multidisciplinary aspects of processing such data and the interplay between clinical, nursing, medical sciences, language technology, computational linguistics, natural language processing/artificial intelligence (NLP/AI), and computer science. The workshop serves as a catalyst for discussing several ongoing research questions that drive both current and future research endeavours, by bringing together researchers from diverse communities. The workshop invited papers describing original research, preferably presenting substantial and completed work, while also welcoming contributions such as negative results, interesting application nuggets, software packages / tools / platforms, small works, or works in progress. It stimulated discussions on various ongoing research questions and challenges by uniting researchers from different communities. We extend our gratitude to the members of the Scientific Program Committee (SPC) for their diligent efforts in reviewing and evaluating all submissions. Each submission received between 2 to 4 reviews, aiding authors in revising and improving their papers accordingly.

There were 12 contributions accepted for the workshop.

Keynote speakers of RaPID-6@MENTAL.ai were:

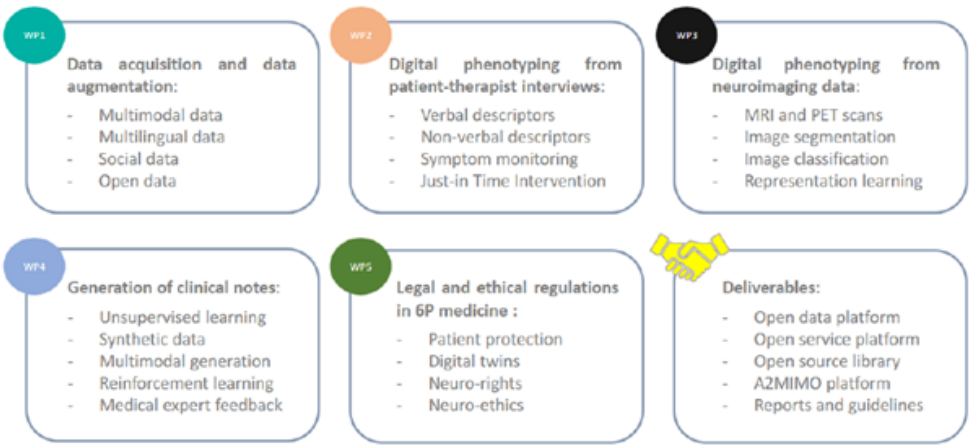
- **Brian MacWhinney**, Teresa Heinz Professor of Cognitive Psychology, Carnegie Mellon University, USA: *TalkBank resources for studying functional communication in language disorders - a survey*; and,
- **Sunny X. Tang**, Associated professor of psychiatry, MD, Feinstein Institutes for Medical Research, Northwell Health, USA: *Speech and language markers for clinical applications in psychiatric disorders*.

Workshop URL: https://spraakbanken.gu.se/en/rapid-2026.

The MENTAL.ai - Summary and Objectives

The MENTAL.AI (actually the MENTAL.AI@CaeSAR*) project aims to develop new AI–based approaches to improve the detection, monitoring, diagnosis, and management of psychiatric disorders and neurodegenerative diseases, such as depression, schizophrenia, and Alzheimer's disease. These conditions currently represent a major public health challenge, affecting millions of people worldwide and having a significant impact on patients, their families, and healthcare systems. The project's goal is to contribute to the development of the 6P medicine (personalized, predictive, preventive, participatory, proof-based, and pathway–oriented) by leveraging the advanced analytical capabilities of AI. By combining various data sources, including text, speech, vision, neuroimaging, behavioral and environmental signals, the project seeks to identify digital markers (digital phenotyping) that allow for earlier symptom detection and a better understanding of disease progression.
The project is based on several complementary research axes. The first involves creating a multimodal and multilingual dataset from clinical interviews conducted via a dedicated digital platform A2MIMO**, allowing for the collection of textual, speech, and behavioral and environmental information (Working Package, WP1). The second aims to develop methods for automatically analyzing patient-therapist interactions to identify linguistic, emotional, or cognitive markers characteristic of specific disorders (WP2). A third axis focuses on analyzing neuroimaging data (MRI, PET) combined with structured tabular information using advanced machine learning techniques to improve diagnostic accuracy, predict disease progression and evaluate patient risks (WP3). The project also explores the use of generative language models to automatically produce clinical reports from medical interviews or imaging data, thereby reducing the administrative burden on healthcare professionals (WP4). Finally, special attention is given to the ethical and legal issues related to the use of AI in the sensitive field of mental health, particularly data protection, patient consent and autonomy, the respect of patients' psychological integrity, liability and accountability, and transparency and disclosure (WP5). MENTAL.AI@CaeSAR relies on an international, interdisciplinary consortium bringing together researchers in computer science, neuroscience, psychiatry, linguistics, cognitive science, and law. By combining these areas of expertise, the project aims to lay the foundation for a new generation of digital tools capable of assisting clinicians in decision-making and improving patient care. In the long term, this work could reshape practices in mental and brain health, supporting earlier diagnoses, precision treatments, and patient-centered monitoring that protects essential neuro-rights.

Start and end dates: 12/2025 - 12/2029
Principal Investigator: Gaël Dias, GREYC UMR 6072



$*$ *CaeSAR: Caen, Stratégie pour l'Accélération en Recherche*
$**$ *A2MIMO: Artificial Agent for Mind Monitoring*

Topics of interest for the RaPID-6@MENTAL.ai

The topics of interest for the workshop session included but were *not* limited to:

- Guidelines, methods and protocols for (remote) data collection and/or annotation (schemas, tools)
- Infrastructure for the domain: building, adapting and sharing of linguistic resources, data sets and tools
- Acquisition and combination of novel data samples; including digital biomarkers, continuous streaming, monitoring and aggregation of measurements; as well as self-reported behavioral and/or physiological and activity data
- Addressing the challenges of representation, including dealing with data sparsity and dimensionality issues, feature combination from different sources and modalities
- Domain adaptation of NLP/AI tools
- Acoustic/phonetic/phonologic, syntactic, semantic, pragmatic and discourse analysis of data; including modeling of perception (e.g. eye-movement measures of reading) and production processes (e.g. recording of the writing process by means of digital pens, keystroke logging etc.); use of gestures accompanying speech and non-linguistic behavior
- Use of wearable, vision, and ambient sensors or their fusion for detection of cognitive disabilities or decline
- (Novel) Modeling and deep / machine learning approaches for early diagnostics, (severity) prediction, monitoring, classification, such as:
  - multimodal learning
  - large pre-trained Transformer language models [LLMs]
  - explainable and interpretable AI models
- Evaluation of the significance of features for screening and diagnostics
- Evaluation of tools, systems, components, metrics, applications and technologies including methodologies making use of NLP/AI; e.g. for predicting clinical scores from (linguistic and/or digital) features
- Digital platforms/technologies for cognitive assessment and brain training
- Evaluation, comparison and critical assessment of resources
- Involvement of medical/clinical professionals and patients
- Ethical, gender bias, legal and safety questions in research with human data in the domain, and how they can be handled
- Deployment, assessment platforms and services as well as innovative mining approaches that can be translated to practical/clinical applications
- Experiences, lessons learned and the future of NLP/AI in the area

Organizing Committee

- Dimitrios Kokkinakis, University of Gothenburg, Sweden *(Workshop's chair)*
- Charalambos Themistocleous, University of Oslo, Norway *(Workshop's co-chair)*
- Gaël Dias, University of Caen Normandie, France
- Kathleen C. Fraser, University of Ottawa, Canada
- Fredrik Öhman, University of Gothenburg, Sweden
- Sebastião Pais, University of Beira Interior, Portugal

Scientific Programme Committee (in alphabetic order)

- Chiara Barattieri di San Pietro, Scuola Universitaria Superiore IUSS di Pavia, Italy
- Patrice Bellot, Aix-Marseille University, France
- Visar Berisha, Arizona State University, USA
- Eric Bui, Caen University Hospital, France
- Gaël Chételat, INSERM, France
- Sunghye Cho, University of Pennsylvania, USA
- Corinne Fredouille, Avignon University, France
- Valantis Fyndanis, University of Technology, Cyprus and University of Oslo, Norway
- Gloria Gagliardi, University of Bologna, Italy
- Natalia Grabar, CNRS, France
- Martínez-Nicolás Israel, Universidad de Salamanca, Spain
- Sandra Anna Just, UiT – The Arctic University, Norway
- Alexandra König, Ki-Elements, Germany
- Klervi Le Dortz, University of Caen Normandie, France
- Alice Lee, University College Cork, Ireland
- Mark Lee, University of Birmingham, UK
- Kristina Lundholm Fors, University of Lund, Sweden
- Patricia Martín-Rodilla, Spanish National Research Council, Spain
- Fabrice Maurel, University of Caen Normandie, France
- Jérémie Pantin, University of Caen Normandie, France
- Javier Parapar, University of La Coruña, Spain
- Alexandre Pauchet, INSA Rouen Normandy, France

- Emily Prud'hommeaux, Boston College, USA
- Masoud Rouhizadeh, University of Florida, USA
- Lina Rydén, University of Gothenburg, Sweden
- Sriparna Saha, Indian Institute of Technology Patna, India
- Johan Skoog, University of Gothenburg, Sweden
- Athanasios Tsanas, University of Edinburgh, UK
- Spyridoula Varlokosta, National and Kapodistrian University of Athens, Greece
- Yaru Wu, University of Caen Normandie, France
- Yasunori Yamada, Boston Medical Science Co., USA

Table of Contents

RaPID-6@MENTAL.ai Workshop Program

Tuesday, May 12, 2026

14:00–16:00 **Session A**
Chair: Dimitrios Kokkinakis

14:00–14:05 **Welcome and Introduction**

14:05–14:50 **Keynote Speaker 1: Brian MacWhinney, Teresa Heinz Professor of Cognitive Psychology, Carnegie Mellon University, USA: "TalkBank resources for studying functional communication in language disorders" (online)**

14:55–16:00 **Oral Session**

14:55–15:15 *Multilingual Cognitive Impairment Detection in the Era of Foundation Models*
Damar Hoogland, Boshko Koloski, Jaya Caporusso, Tine Kolenik, Senja Pollak, Christina Manouilidou and Matthew Purver

15:15–15:35 *The Icelandic Language Biobank: Data Collection through a Clinical Analysis Platform*
Iris Nowenstein, Naizeth Núñez Macías, Gunnar Thor Örnólfsson, Stefán Ólafsson, Bryndís Bergþórsdóttir, Iðunn Kristínardóttir and Hinrik Hafsteinsson

15:35–15:55 **Invited Speaker: Professor Gaël Dias, Université de Caen Normandie, France: "MENTAL.ai - Artificial Intelligence for Mental and Brain Health"**

15:55–16:00 **Questions or comments to the sessions's presenters**

16:00–16:30 ***Afternoon Coffee Break***

Tuesday, May 12, 2026 (continued)

16:00–18:15 **Session B: Poster and Oral Sessions**
Chair: Charalambos Themistocleous

16:00–18:20 **Poster Session** (Note: The poster session will begin during the coffee break. Poster presenters are expected to be available at their posters until 17:00, while the posters will remain on display until the end of the workshop)

Disfluencies and ASR Performance on Swedish Spontaneous Speech from the 'Trip to Stockholm' Discourse Narrative Task
Dimitrios Kokkinakis, Herbert Lange and Ricardo Muñoz Sánchez

ALBA: An Automated Framework for Benchmarking Clinical Language Biomarkers against Standardized Corpora
Charalambos Themistocleous and Brielle C. Stark

On Automatic Detection of Cognitive Decline
Fabio Tamburini

Benchmarking NLP-supported Language Sample Analysis for Swiss Children's Speech
Anja Ryser, Yingqiang Gao and Sarah Ebling

Resource-Efficient LLMs for Depression Symptoms Screening: Performance and Limitations in Zero Shot Setting
Muhammad Rizwan and Jure Demšar

CNSocialDepress: A Chinese Social Media Dataset for Depression Risk Detection and Structured Analysis
Jinyuan Xu, Tian Lan, Xintao Yu, Xue He, Hezhi Zhang, Ying Wang, Mathieu Valette, Pierre Magistry and Lei Li

Depression Detection in Modern Greek
Vivian Stamou, George Mikros, George Markopoulos and Spyridoula Varlokosta

Profiling Psychopathic Behavior Using Machine Learning
Avi Treistman, Tehilla David, Sivan Levi and Dror Mughaz

Tuesday, May 12, 2026 (continued)

17:00–17:40 **Keynote Speaker 2: Sunny X. Tang, M.D., Assistant Professor, Department of Psychiatry, Feinstein Institutes for Medical Research / Zucker Hillside Hospital, Glen Oaks, NY, USA: "Speech and Language Markers for Clinical Applications in Psychiatric Disorders" (online)**

17:40–18:20 **Oral Session**

17:40–17:55 *Developing Annotation Guidelines for CSAM Prevention Interventions: Psychosocial Risk and Protective Factors Grounded in Research and Clinical Practice*
Vera Czehmann, Christine Hovhannisyan, Lena Elisabeth Hoffmann, Paula Busch, Ibrahim Baroud, Sebastian Möller, Roland Roller, Hannes Gieseler and Lisa Raithel

17:55–18:10 *Automatic Detection of Direct and Self-Repetitions in Naturalistic Speech Recordings of French- and Dutch-Speaking Autistic Children*
Federica Beccaria, Marie Kolenberg, Pierre Labendzki, Inge Zink and Mikhail Kissine

18:10–18:15 **Questions or comments to the sessions's presenters**

18:15–18:20 **Conclusions**

Multilingual Cognitive Impairment Detection in the Era of Foundation Models

Damar Hoogland[1] **Boshko Koloski**[1,2] **Jaya Caporusso**[1,2] **Tine Kolenik**[3]
Ana Zwitter Vitez[4] **Senja Pollak**[1] **Christina Manouilidou**[4] **Matthew Purver**[1,5]
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] Institute of Synergetics and Psychotherapy Research, Paracelsus Medical University, Salzburg, Austria
[4] University of Ljubljana, Ljubljana, Slovenia
[5] Queen Mary University of London, London, UK

Abstract

We evaluate cognitive impairment (CI) classification from transcripts of speech in English, Slovene, and Korean. We compare zero-shot large language models (LLMs) used as direct classifiers under three input settings—transcript-only, linguistic-features-only, and combined—with supervised tabular approaches trained under a leave-one-out protocol. The tabular models operate on engineered linguistic features, transcript embeddings, and early or late fusion of both modalities. Across languages, zero-shot LLMs provide competitive no-training baselines, but supervised tabular models generally perform better, particularly when engineered linguistic features are included and combined with embeddings. Few-shot experiments focusing on embeddings indicate that the value of limited supervision is language-dependent, with some languages benefiting substantially from additional labelled examples while others remain constrained without richer feature representations. Overall, the results suggest that, in small-data CI detection, structured linguistic signals and simple fusion-based classifiers remain strong and reliable signals.

Keywords: cognitive decline detection, large language models, tabular foundation models, feature fusion

1. Introduction

Cognitive impairment (CI) refers to a state in which a person's cognitive functioning is below the expected level and is a diagnosable condition (Ray and Davidson, 2014). CI can involve varying degrees of deterioration of cognitive abilities such as memory, attention, executive functioning, and language, and it is often associated with neurodegenerative diseases like Alzheimer's disease (AD) and other conditions that cause dementia (Morley, 2018). Although relatively advanced CI associated with such diseases is often preceded by a mild cognitive impairment (MCI) phase, not all individuals with MCI progress to dementia (Petersen, 2016). Early identification of CI is essential to enable timely and appropriate clinical intervention, patient support, and participation in preventive or therapeutic programmes (Livingston et al., 2024).

Traditional diagnostic assessments for cognitive impairment include neurophysiological tests (Nasreddine et al., 2005), clinical and functional assessments (O'Bryant et al., 2008), neuroimaging and biomarker assessments (Hampel et al., 2018), and clinical interviews and observation (McKhann et al., 2011). Many of these assessments involve language, as individuals with CI frequently exhibit lexical retrieval difficulties, semantic degradation, syntactic simplification, and reduced discourse organisation, reflecting underlying deterioration in semantic memory and executive control (Taler and Phillips, 2008; Fraser et al., 2015; Boschi et al., 2017). For example, picture description tasks such as the Cookie Theft task (Goodglass and Kaplan, 1983) prompt individuals to describe a complex visual scene. The resulting descriptions enable qualitative and quantitative assessment of language production, including lexical retrieval, syntactic formulation, fluency, informativeness, and narrative organisation.

However, traditional diagnostic tools can be limited in their ability to detect early cognitive changes (Trzepacz et al., 2015) and often require in-person administration by trained professionals, making them resource-intensive, time-consuming, and impractical for frequent or large-scale screening (Slegers et al., 2018). Their outcomes can furthermore be affected by education and language background, introducing cultural and linguistic bias (Ramos-Henderson et al., 2025). Finally, because they are typically administered infrequently in clinical settings, they provide only snapshots of cognition rather than a continuous measure of change (Patnode et al., 2020).

Computational methods—particularly those leveraging Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL)—address several limitations of traditional assessments by enabling automated and fine-grained analysis of spontaneous speech. Such approaches can capture subtle linguistic and discourse changes that may precede clinical diagnosis, operate remotely and non-invasively, and allow for repeated or continuous monitoring over time (De la Fuente Garcia et al., 2020). While computational approaches risk inheriting or

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 1–12
12 May 2026.

amplifying linguistic and cultural biases present in existing assessments, they also have the potential to support diagnosis by trained clinicians when developed and evaluated responsibly.

With recent advances in pretrained foundation models and increased accessibility of computational resources, research on automatic CI detection from language has begun shifting from classical ML approaches that rely on expert-selected symbolic features towards methods built around pretrained language models and general-purpose tabular predictors (reviewed in Section 2). In this study, we compare inference-only foundation-model baselines (multilingual Large Language Models (LLMs) and tabular foundation models) against classical ML and fusion-based approaches across three languages, under unified zero-shot, few-shot, and leave-one-out evaluation protocols.

2. Related Work

Many studies investigating the detection and prediction of cognitive decline have employed classical ML approaches (Huang et al., 2024; Kaser et al., 2024). For example, Luz et al. (2021) employed a range of classical ML models—including linear discriminant analysis, decision trees, k-nearest neighbours, random forests, and support vector machines—to classify Alzheimer's vs. healthy speech and to predict scores obtained with the Mini-Mental State Examination test. These models were built on manually engineered acoustic and linguistic feature sets.

More recently, others have moved to using Large Language Models (LLMs) for feature extraction. For example, de Arriba-Pérez et al. (2024) automatically extracted a large set of high-level, content-independent linguistic features from free dialogues using ChatGPT[1] prompts, alongside traditional n-gram features. These features were then analysed, selected, and used to train classical ML classifiers to detect cognitive decline. Other studies employed DL techniques, such as BERT-based transformer classifiers (Mao et al., 2022; Ilias and Askounis, 2022; Pahar et al., 2025; Zhu et al., 2022).

Large Language Models (LLMs) are increasingly employed as direct classifiers for cognitive decline detection (e.g., Jiang et al., 2026). Zheng et al. (2024) used pre-trained LLaMA-2[2] models with prompt engineering, Low-Rank Adaptation (LoRA) fine-tuning, and conditional learning to classify AD from speech transcripts of the ADReSS dataset (Luz et al., 2020). Guan et al. (2025) presented CD-Tron, a system built on the clinical LLM GatorTron[3], fine-tuned on labelled electronic health record note sections to detect early cognitive decline, and reported substantial improvements over smaller transformers and GPT-4. Botelho et al. (2024) used LLMs both as predictors and as feature extractors, prompting LLMs to produce interpretable macro-descriptors (e.g., coherence, lexical diversity, word-finding difficulty) which were then used as inputs to simple classifiers for AD detection.

Most studies reviewed above, both classical ML and LLM-based approaches, have focused on monolingual settings, with the majority using only English data. This makes it difficult to assess how well these paradigms transfer across languages and elicitation paradigms.

Tabular foundation models have recently been proposed as general-purpose predictors for structured data, enabling strong performance on small tabular datasets via in-context learning and reducing the need for task-specific training (Hollmann et al., 2025). In the cognitive decline domain, Ding et al. (2026) apply TabPFN to longitudinal AD modelling on the TADPOLE benchmark, predicting clinical diagnosis and cognitive scores from tabular patient records. Related work explores TabPFN-based approaches for dementia-related prediction in other neurodegenerative settings, such as predicting Parkinson's disease dementia using a hybrid LightGBM–TabPFN model with SHAP-based interpretability (Tran and Byeon, 2024). These studies do not use language data directly, and comparisons to language-based CI detection remain limited.

2.1. This study

Prior work has not systematically compared (i) classical ML models with expert-assisted linguistic features, (ii) embedding-based tabular models, (iii) tabular foundation models, and (iv) prompted LLM classifiers under a unified protocol and across multiple languages. In the present study, we conduct within-language experiments for three languages—English, Slovene, and Korean—and compare symbolic-feature-based models, embedding-based models, fusion strategies, and zero-shot LLM baselines under unified leave-one-out, zero-shot, and few-shot protocols. We further analyse representation alignment between symbolic features and embeddings to understand when and why multimodal fusion helps in small-data CI detection.

Our research questions (RQs) are as follows.

- **RQ1**: How well do LLMs (specifically, gpt-oss-20b and med-gemma-27b) discriminate CI vs. Healthy Control (HC) participants across three languages under zero-shot prompting?
- **RQ2**: How sensitive to the input modality (transcript-only, linguistic-features-only, or transcript+features) is the performance of LLMs?

[1] https://chat.openai.com/

[2] https://www.llama.com/llama2/

[3] https://huggingface.co/UFNLP/gatortron-base

- **RQ3**: Do expert-assisted symbolic linguistic features improve CI classification when paired with tabular models (TabPFN, RealMLP, and classical baselines) compared to (i) embedding-only representations and (ii) LLM-based classifiers, and are gains consistent across languages and evaluation paradigms?
- **RQ4**: Which integration strategy yields the best and most stable performance across languages among embeddings-only, symbolic-features-only, and multimodal fusion (normalised concatenation / feature reweighting, or late fusion), under both leave-one-out and few-shot evaluation?

3. Data and preprocessing

3.1. Datasets

To evaluate cross-linguistic generalisability and performance stability, we ran parallel experiments on three languages: English, Slovene, and Korean. The English and Slovene datasets were obtained from corpora of recordings of picture description tasks, while the Korean data came from a corpus of structured interviews. The English and Slovene datasets include participants with AD and HCs, while the Korean dataset includes participants with MCI and HCs. In all experiments we treat the positive class as PATIENT (AD or MCI, depending on the dataset) and the negative class as CONTROL (HC).

The included datasets are listed below, and Table 1 summarises the number of participants and diagnostic labels per dataset.

English: The English dataset consists of a subset of Cookie Theft Picture Descriptions from the Pitt Corpus (Becker et al., 1994; the original corpus is available on DementiaBank, MacWhinney et al., 2011), pre-processed for the ADReSS challenge (Luz et al., 2020). It includes participants with AD and Healthy Controls (HCs).

Slovene: The Slovene dataset was collected as part of the CogLiTreat project, which investigated behavioural and transcranial magnetic stimulation interventions for language disorders. It includes recordings and transcripts of Slovene AD patients and control participants from the Ljubljana region who described the New Cookie Theft picture (Berube et al., 2019). The participants' responses were recorded and the detection of speech and silence was performed automatically using Praat (Boersma and Weenink, 2021). The recordings were orthographically transcribed by one of the interviewers and cross-checked by an independent native speaker of Slovene. Finally, each utterance was assigned to the participant or interviewer manually by the first author of the present study. We note two potential confounding factors. First, the patient recordings were delivered in a different file format (m4a) than the control group (WAV), which may have introduced confounds during preprocessing (e.g., silence detection may behave differently across formats). Second, the experimenter differed between the two groups, which may affect language use due to conversational alignment effects (Pickering and Garrod, 2004; Freud et al., 2018). We return to these issues in Section 8.

Korean: The Korean dataset was obtained from the Kang corpus, available on DementiaBank (MacWhinney et al., 2011). The Kang corpus includes participants with MCI and HCs. Each participant took part in a structured interview consisting of 16 questions. The corpus includes manual transcriptions.

3.2. Transcript pre-processing

For each dataset, we extracted the participant utterances and removed non-orthographic annotations (e.g., the use of '(.)' to indicate short pauses in the Pitt corpus).

For the English and Slovene datasets, each utterance was processed using that language's model from the Stanza NLP library (Qi et al., 2020). After tokenisation, tokens labelled as punctuation were excluded, and for each remaining token we extracted the surface form, lemma, universal part-of-speech (UPOS) tag, dependency relation, and syntactic head index.

For Korean, we used the MeCab morphological analyser for tokenisation and part-of-speech (POS) tagging (Kudo, 2005). Tokens labelled as punctuation were excluded (SF: sentence-final punctuation; SP: comma/pause; SS: brackets/quotation marks; SE: ellipsis; SO: other symbols), and the remaining POS tags were converted to their UPOS equivalents (Park and Tyers, 2019). Dependency-based features were not extracted for Korean, as MeCab does not provide dependency parsing.

3.3. Features

From the preprocessed participant-only transcripts we extracted eleven linguistic features that were reported as indicative of AD-related language change in a recent systematic review (Shankar et al., 2025). All features were calculated over all utterances per participant and per task. For Korean, two

Dataset	Condition	Patients	Controls	Total	Patient/Control (%)
English	AD	78	78	156	50.0 / 50.0
Slovene	AD	12	15	27	44.4 / 55.6
Korean	MCI	40	37	77	51.9 / 48.1
Total	–	**130**	**130**	**260**	**50.0 / 50.0**

Table 1: Participant statistics per dataset with within-dataset class proportions. AD: Alzheimer's Disease; MCI: Mild Cognitive Impairment.

dependency-based features (idea density and syntactic complexity) could not be computed (see Section 3.2); these values are treated as missing and handled by the imputation procedure described in Section 4.2.

Speech Rate The number of words uttered by the participant, divided by the total duration of the task in seconds. In English and Slovene, we divided the number of words by the duration from the start of the first participant utterance to the end of the last participant utterance, without excluding interviewer speech. In Korean, we excluded interviewer speech from the duration because the interviewer took a more active role due to the structured nature of the task. Interviewer speech could not be consistently removed from the total duration in English because accurate time-stamps were not available.

Type-Token Ratio The number of unique tokens in the participant's speech divided by the total number of words they uttered.

Repetitiveness The mean cosine distance between embeddings (produced by Sentence-BERT; Reimers and Gurevych, 2019) of each consecutive pair of participant utterances.

Coherence The mean cosine distance of embeddings (produced by Sentence-BERT; Reimers and Gurevych, 2019) between all possible pairs of different participant utterances.

Familiarity The mean familiarity per unique word used by the participant. Familiarity is expressed as the number of occurrences per million words in speech corpora, provided by frequency reference lists for each language (Dobrovoljc, 2018; Leech et al., 2014; Kim et al., 2024).

Idea Density The number of main verbs divided by the total number of tokens uttered by the participant. Not computed for Korean (Section 3.2).

Syntactic Complexity The mean maximal syntactic tree depth per participant utterance. Not computed for Korean (Section 3.2).

Verb Ratio The number of verbs divided by the total number of tokens.

Noun Ratio The number of nouns divided by the total number of tokens.

Pronoun Ratio The number of pronouns divided by the total number of tokens.

Pronoun to Noun Ratio The number of pronouns divided by the total number of nouns.

4. Modelling Methodology

4.1. Task and Data

We study binary classification of CI (AD or MCI, depending on the dataset) versus HC from speech-derived inputs. We evaluate performance separately for three languages: English, Slovene, and Korean (Section 3.1). All experiments are conducted within-language (i.e., training and evaluation never mix languages), and results are reported per language and aggregated across languages.

4.2. Input Representations

We compare three input configurations derived from each sample.

(1) Symbolic linguistic features. We use an 11-dimensional vector of expert-assisted textual features:

$$\mathbf{x}_{\text{feat}} \in \mathbb{R}^{11},$$

corresponding to the features listed in Section 3.3. For Korean, two feature dimensions are missing and are handled by imputation within each fold (see below).

(2) Embedding-based representation. We compute a fixed-dimensional dense embedding from the transcript of the participant's utterances using a frozen multilingual embedding model (`google/embedding-gemma-300m`):

$$\mathbf{x}_{\text{emb}} = f_{\text{emb}}(t), \qquad \mathbf{x}_{\text{emb}} \in \mathbb{R}^{d}.$$

Embeddings are computed once and reused across all evaluation runs.

(3) Fusion of embeddings and features. We consider two fusion strategies that combine the embedding and symbolic feature modalities.

For *early fusion*, we preprocess each modality independently, re-weight the symbolic features to compensate for the dimensionality imbalance ($d \gg 11$), and concatenate into a single vector passed to one classifier:

$$w = \sqrt{\frac{d}{11}}, \qquad \mathbf{x}_{\text{early}} = \left[\, \tilde{\mathbf{x}}_{\text{emb}} \;;\; w \cdot \tilde{\mathbf{x}}_{\text{feat}} \,\right].$$

For *late fusion*, we train two independent classifiers of the same family—one on $\tilde{\mathbf{x}}_{\text{emb}}$ and one on $\tilde{\mathbf{x}}_{\text{feat}}$—and combine their outputs by averaging the predicted class probabilities:

$$\hat{p}_{\text{late}} = \frac{1}{2}\Big(\hat{p}_{\text{emb}} + \hat{p}_{\text{feat}}\Big), \qquad \hat{y}_{\text{late}} = \mathbf{1}\big[\hat{p}_{\text{late}} \geq 0.5\big].$$

This decision-level combination allows each modality to be preprocessed and modelled independently before fusion.

Preprocessing. To prevent data leakage, all preprocessing steps are fit using training data only within each evaluation fold or episode. We apply median imputation per feature dimension using a `SimpleImputer` fit on the training split. We standardise each feature dimension using z-score normalisation (`StandardScaler`) fit on the training split. For both fusion variants, embeddings and symbolic features are imputed and standardised separately.

4.3. Models

We compare tabular foundation models, classical ML baselines, and prompted LLMs.

4.3.1. Tabular and Foundational Models

We train the following classifiers per language and representation (embeddings, features, early fusion, and late fusion): TabPFN (foundational in-context tabular classifier), RealMLP (tabular deep learning baseline), Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM) with both linear (SVM-Linear) and radial basis function (SVM-RBF) kernels, LightGBM (LGBM), and k-nearest neighbours (k-NN) with $k \in \{3, 5, 7\}$. For late fusion, each model family is instantiated twice per fold or episode (once per modality) and their predicted probabilities are averaged as described in Section 4.2.

4.3.2. LLM-based Classification

We evaluate two LLMs as direct classifiers: gpt-oss-20b and med-gemma-27b. Both are used in an inference-only setting via an OpenAI-compatible API endpoint served through vLLM.

We test three prompt variants: (1) transcript-only, (2) linguistic-only (the 11 symbolic features rendered as a numeric list), and (3) full-data (the transcript concatenated with the symbolic feature list). Full prompt templates are provided in Appendix 1.

Models are instructed to output exactly one token: CONTROL or PATIENT. Where supported by the inference server, we enforce a guided-choice constraint restricting outputs to {CONTROL, PATIENT}. Temperature is set to 0 for deterministic decoding unless otherwise stated.

4.4. Evaluation Protocols

All evaluations are conducted *within-language*. We report per-language performance and aggregated averages across languages. To ensure direct comparability between tabular models and LLMs across all evaluation settings, both model families share the same outer leave-one-out loop and the same episodic sampling scheme for few-shot conditions.

4.4.1. Full-Data Evaluation (Leave-One-Out)

For each language L, we perform leave-one-out cross-validation over all n_L samples. Each fold holds out exactly one sample i for testing and uses the remaining $n_L - 1$ samples as the training pool. For tabular models, preprocessing is fit on the training pool only and applied to both train and test. For LLMs, no demonstrations are used in the zero-shot condition. We refer to this setting as the *full-data* evaluation. For each language, we also report a majority-class predictor trained and evaluated under the same LOO protocol.

4.4.2. Few-Shot Episodic Evaluation

To study few-shot behaviour under a unified and directly comparable protocol, we apply the same outer LOO loop and inner episodic sampling scheme to tabular models.

Table 2: LOO Macro-F1 across all models and input configurations. Classical ML, tabular foundation models (TFMs), gradient boosting, and LLM zero-shot baselines. Best per language in **bold**. † = zero-shot (no training data).

Method	Input	English	Slovene	Korean
Non-learning baseline				
Majority	—	0.333	0.357	0.342
LLMs (zero-shot†)				
MedGemma-27B†	Full	0.361	0.518	0.595
	Ling.	0.333	0.357	0.342
	Trans.	0.413	0.492	0.579
GPT-OSS-20B†	Full	0.413	0.617	0.609
	Ling.	0.500	0.555	0.349
	Trans.	0.556	0.603	0.621
Tabular foundation models				
TabPFN	Emb	0.343	**0.852**	0.523
	Feat	0.814	0.454	0.740
	Early	0.784	**0.852**	0.792
	Late	**0.816**	0.727	0.753
RealMLP	Emb	0.493	0.802	0.566
	Feat	0.746	0.682	0.692
	Early	0.743	0.701	0.737
	Late	0.738	0.754	0.680
Gradient boosting				
LGBM	Emb	0.549	0.357	0.342
	Feat	0.782	0.357	0.621
	Early	0.763	0.357	0.621
	Late	0.776	0.357	0.633
Classical ML				
LR	Emb	0.549	**0.852**	0.523
	Feat	0.807	0.540	0.739
	Early	0.801	0.735	0.792
	Late	0.788	0.727	**0.805**
RF	Emb	0.549	**0.852**	0.523
	Feat	0.781	0.635	0.675
	Early	0.665	**0.852**	0.714
	Late	0.737	**0.852**	0.714
SVM (linear)	Emb	0.549	**0.852**	0.523
	Feat	0.813	0.540	0.727
	Early	0.750	0.659	0.778
	Late	0.769	0.814	0.789
SVM (RBF)	Emb	0.549	**0.852**	0.523
	Feat	0.806	0.442	0.714
	Early	0.738	**0.852**	0.766
	Late	0.807	**0.852**	0.712
kNN-3	Emb	0.333	0.357	0.523
	Feat	0.712	0.508	0.726
	Early	0.654	**0.852**	0.778
	Late	0.698	0.250	0.476
kNN-5	Emb	0.325	0.308	0.523
	Feat	0.750	0.463	0.674
	Early	0.652	0.814	0.779
	Late	0.582	0.583	0.476
kNN-7	Emb	0.410	**0.852**	0.523
	Feat	0.762	0.365	0.673
	Early	0.665	**0.852**	0.752
	Late	0.543	0.735	0.500

For each held-out test sample i, we sample k examples per class uniformly at random from the remaining $n_L - 1$ samples (the support set), using only these $2k$ samples for training. We repeat this sampling for $E = 3$ episodes with different random seeds and aggregate predictions across episodes by majority vote; for models producing calibrated probabilities, we additionally average scores before thresholding. We evaluate $k \in \{1, 2, 3, 5\}$. For tabular models, the $2k$ support samples are used to fit the classifier (with preprocessing fit on the same $2k$ samples). For late fusion, two classifiers are fit on the $2k$ samples (one per modality) and their probabilities are averaged.

4.4.3. Zero-Shot Evaluation for LLMs

For each language and prompt modality, we classify each sample independently with no labeled demonstrations.

Evaluation For each language, model, and evaluation mode we report Macro-F1. To estimate variability due to episodic sampling and any stochasticity in model training, we repeat the full tabular evaluation pipeline three times with different random seeds. Results are aggregated by Language, Model, Evaluation Mode and reported as mean. LLM evaluations are deterministic at temperature 0 and are therefore reported as single-run results.

5. Results and Discussion

Leave-one-out (LOO) Macro-F1 results are reported in Table 2. Few-shot results (embeddings-only; k shots per class) are reported in Table 3. We discuss the findings by research question.

Table 3: Few-shot Macro-F1 (embeddings-only, k shots/class, 3 seeds) vs. LLM zero-shot reference. Best overall per language in **bold**.

Method	k	English	Slovene	Korean
LLM zero-shot reference (best across models & modalities)				
Best LLM†	0	**0.556**	0.617	0.621
Tabular foundation models				
TabPFN	1	0.439	0.846	0.538
	2	0.442	0.815	0.504
	3	0.436	**0.852**	0.560
	5	0.472	**0.852**	0.532
RealMLP	1	0.404	**0.852**	0.497
	2	0.461	0.815	0.522
	3	0.448	0.839	0.468
	5	0.455	**0.852**	0.570
Classical ML				
LR	1	0.400	**0.852**	0.497
	2	0.431	**0.852**	0.568
	3	0.444	**0.852**	0.552
	5	0.449	**0.852**	**0.667**
RF	1	0.382	**0.852**	0.497
	2	0.462	**0.852**	0.568
	3	0.458	**0.852**	0.552
	5	0.482	**0.852**	0.452

RQ1: LLM zero-shot CI detection. Zero-shot LLM performance is generally limited relative to supervised tabular models. The best zero-shot LLM result is GPT-OSS-20B on Korean (0.621 Macro-F1), which is +0.279 above the majority baseline (0.342). On English and Slovene, LLM performance varies substantially by prompt modality; for example, MedGemma-27B collapses to majority-class behaviour in the linguistic-only setting on English (0.333 Macro-F1). Despite being a medical-domain model, MedGemma-27B underperforms GPT-OSS-20B across all three languages, suggesting general instruction-following behaviour and robustness to prompt formatting may matter more than domain specialisation in this setting.

RQ2: Modality sensitivity. No single input modality consistently dominates across languages and models. Transcript-only input works best for English and Korean with GPT-OSS-20B, while full-data prompts (transcript + features) do not reliably outperform single-modality prompts. This suggests that, in an inference-only setting, LLMs may struggle to integrate heterogeneous numeric and textual evidence as reliably as simpler tabular fusion approaches.

RQ3: Symbolic features + tabular models vs. LLMs. Tabular models with symbolic features decisively outperform zero-shot LLM baselines (+0.18 to +0.26 Macro-F1 across languages when comparing best results per language). The 11-feature vector is highly informative: on English, TabPFN achieves 0.814 with features alone, a +0.471 improvement over embeddings-only (0.343). Classical ML baselines remain competitive, with LR achieving the best Korean result (0.805, late fusion). Slovene is an exception where embeddings dominate features (0.852 vs. 0.454 for TabPFN), though potential confounds (different recording formats and experimenters across groups) may inflate embedding-based separability (Sections 3.1 and 8).

RQ4: Fusion strategies. Early fusion generally performs best for Slovene and yields strong performance for Korean, indicating that combining complementary information sources can improve stability across datasets. Features-only is strongest for English, where the engineered linguistic signal is particularly predictive and adding high-dimensional embeddings can dilute that signal for some models. To better understand when fusion helps, we analyse alignment between feature space and embedding space (Table 4). English and Korean show near-orthogonal representations (CKA 0.024 and 0.016; low Overlap@5), consistent with fusion providing complementary information. Slovene exhibits higher alignment (CKA 0.181; Overlap@5 0.200), suggesting that in this dataset embeddings may already encode much of the feature-level signal (or reflect dataset-specific confounds).

Few-shot vs. zero-shot LLMs. Even with minimal labels, tabular models can match or exceed LLM zero-shot performance for some languages. For Slovene, embedding-based few-shot models exceed the best LLM from just k=1 example per class (0.846–0.852 vs. 0.617). For Korean, LR reaches 0.667 at k=5, exceeding the best LLM (0.621). In English, the best LLM zero-shot result (0.556) remains higher than embedding-only few-shot baselines, highlighting the importance of incorporating symbolic features and/or stronger representations for small-data supervision.

Table 4: Feature–embedding space alignment. Low CKA and Spearman ρ indicate weak alignment between representations; Procrustes disparity (Williams et al., 2021) near 2.0 indicates geometric dissimilarity. Overlap@5 = shared kNN neighbours; Purity@5 = same-class neighbours.

Language	CKA	Spearman ρ	Procrustes	Overlap@5	Purity$_{feat}$@5	Purity$_{emb}$@5
English	0.024	0.001	1.758	0.032	0.658	0.501
Slovene	0.181	0.116	1.432	0.200	0.511	0.563
Korean	0.016	0.005	1.835	0.049	0.610	0.610

6. Conclusion

We evaluated CI detection across three languages (English, Slovene, and Korean) comparing LLM

zero-shot prompting, tabular foundation models, and classical ML. LLMs provide usable no-training baselines (best Macro-F1: 0.621), but supervised tabular models—especially those using expert-assisted symbolic features and fusion—achieve substantially higher performance (+0.18 to +0.26 Macro-F1 over the best LLM per language). Across datasets, lightweight classical models remain highly competitive, with TabPFN reaching 0.816 on English (late fusion) and LR reaching 0.805 on Korean (late fusion). Alignment analysis indicates that feature and embedding representations are weakly aligned in English and Korean, providing principled support for fusion; Slovene shows higher alignment, consistent with embeddings dominating in that dataset. For practical deployment in small-data CI detection, tabular models operating on transparent linguistic markers offer a strong and interpretable alternative to inference-only LLM classification, while LLM prompting remains a useful reference point when training data are unavailable.

7. Code Availability

The source code is publicly available at https://github.com/bkolosk1/foundational-ci-detection.

8. Limitations

The Slovene dataset (n=27) exhibits potential confounds: different recording formats and experimenters for patient and control groups may inflate embedding-based performance. Two symbolic features (idea density and syntactic complexity) are unavailable for Korean due to parser limitations, and are therefore treated as missing and imputed; this reduces the amount of available symbolic information for that language. While the symbolic features are individually interpretable, fusion with 768-dimensional embeddings reduces transparency; future work should investigate explanation methods (e.g., SHAP) for fusion models and assess robustness across elicitation paradigms. The relatively small dataset size, even if it is typical for this field, restricts the strength and generalisability of our conclusions. Moreover, interviewer speech was not excluded from the speech-rate calculation in the English and Slovene data, which may have affected these measures. As a result, the reported speech-rate values may not fully reflect participant speech alone and should be interpreted with caution. A further limitation is that the study does not examine potential sources of bias related to participant characteristics such as age, education, and linguistic background, all of which may affect linguistic performance and, in turn, influence classification outcomes. It also does not account for differences in cognitive or brain reserve. Assessing such demographic and individual differences is essential before any clinical deployment. Finally, two of the three datasets use picture description tasks; generalisation to other elicitation paradigms (spontaneous speech, narrative recall) remains to be investigated.

Acknowledgments

We acknowledge the financial support from the Slovenian Research Agency ARIS via the projects Cross-Lingual Analysis for Detection of Cognitive Impairment in Less-Resourced Languages (CroDeCo; J6-60109), and Natural Language Processing for Corpus Analysis in the Medical Humanities (BI-VB/25-27-021), and research core funding for the programme Knowledge Technologies (P2-0103).

BK is funded by the Young Researcher Grant PR-12394, and JC by the Young Researcher Grant PR-13409.

The English dataset was based on the Pitt Corpus (Becker et al., 1994), which was produced with the support of grants NIA AG03705 and AG05133 to the original authors of the corpus.

9. Bibliographical References

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Shauna Berube, Jodi Nonnemacher, Cornelia Demsky, Shenly Glenn, Sadhvi Saxena, Amy Wright, Donna C Tippett, and Argye E Hillis. 2019. Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American Journal of Speech-Language Pathology*, 28(1S):321–329.

Paul Boersma and David Weenink. 2021. Praat: Doing phonetics by computer (version 6.4.06) [computer software]. http://www.praat.org/.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in Psychology*, 8:269.

Catarina Botelho, John Mendonça, Anna Pompili, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2024. Macro-descriptors for alzheimer's

disease detection using large language models. In *Interspeech*, pages 1975–1979.

Francisco de Arriba-Pérez, Silvia García-Méndez, Javier Otero-Mosquera, and Francisco J González-Castaño. 2024. Explainable cognitive decline detection in free dialogues with a machine learning approach based on pre-trained large language models. *arXiv preprint arXiv:2411.02036*.

Sofia De la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

Yilang Ding, Jiawen Ren, Jiaying Lu, Gloria Hyunjung Kwak, Armin Iraji, Shengpu Tang, and Alex Fedorov. 2026. Longitudinal progression prediction of alzheimer's disease with tabular foundation model.

Kaja Dobrovoljc. 2018. Gos corpus n-grams 2.0. https://www.clarin.si/repository/xmlui/handle/11356/1195.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Debora Freud, Ruth Ezrati-Vinacour, and Ofer Amir. 2018. Speech rate adjustment of adults during conversation. *Journal of fluency disorders*, 57:1–10.

Harold Goodglass and Edith Kaplan. 1983. *The assessment of aphasia and related disorders*.

Hao Guan, John Novoa-Laurentiev, and Li Zhou. 2025. Cd-tron: Leveraging large clinical language model for early detection of cognitive decline from electronic health records. *Journal of Biomedical Informatics*, page 104830.

Harald Hampel, Sid E O'Bryant, José L Molinuevo, Henrik Zetterberg, Colin L Masters, Simone Lista, Steven J Kiddle, Richard Batrla, and Kaj Blennow. 2018. Blood-based biomarkers for alzheimer disease: mapping the road to the clinic. *Nature Reviews Neurology*, 14(11):639–652.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

Lihe Huang, Hao Yang, Yiran Che, and Jingjing Yang. 2024. Automatic speech analysis for detecting cognitive decline of older adults. *Frontiers in Public Health*, 12:1417966.

Loukas Ilias and Dimitris Askounis. 2022. Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4153–4164.

Lei Jiang, Yue Zhou, and Natalie Parde. 2026. What do llms know about alzheimer's disease? fine-tuning, probing, and data synthesis for ad detection. *arXiv preprint arXiv:2602.11177*.

Alyssa N Kaser, Laura H Lacritz, Holly R Winiarski, Peru Gabirondo, Jeff Schaffert, Alberto J Coca, Javier Jiménez-Raboso, Tomas Rojo, Carla Zaldua, Iker Honorato, et al. 2024. A novel speech analysis algorithm to detect cognitive impairment in a spanish population. *Frontiers in Neurology*, 15:1342907.

Jin-seo Kim, Anna Seo Gyeong Choi, and Sunghye Cho. 2024. Kofren: Comprehensive korean word frequency norms derived from large scale free speech corpora. In *Proceedings of the Joint Conference on Language Resources and Evaluation*.

Takumitsu Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Routledge.

Gill Livingston, Jonathan Huntley, Kathy Y Liu, Sergi G Costafreda, Geir Selbæk, Suvarna Alladi, David Ames, Sube Banerjee, Alistair Burns, Carol Brayne, et al. 2024. The lancet commissions. *Lancet*, 404:572–628.

Saturnino Luz, Farhana Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *Proceedings of Interspeech 2020*, pages 2172–2176.

Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. 2021. Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science*, 3:780169.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25:1286–1307.

Chengsheng Mao, Jie Xu, Luke Rasmussen, Yikuan Li, Prakash Adekkanattu, Jennifer Pacheco, Borna Bonakdarpour, Robert Vassar, Guoqian Jiang, Fei Wang, et al. 2022. Ad-bert: using pre-trained contextualized embeddings to predict the progression from mild cognitive impairment to alzheimer's disease. *arXiv preprint arXiv:2212.06042*.

Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. 2011. The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269.

John E Morley. 2018. An overview of cognitive impairment. *Clinics in Geriatric Medicine*, 34(4):505–513.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2005. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Sid E O'Bryant, Stephen C Waring, C Munro Cullum, James Hall, Laura Lacritz, Paul J Massman, Philip J Lupo, Joan S Reisch, Rachelle Doody, Texas Alzheimer's Research Consortium, et al. 2008. Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study. *Archives of Neurology*, 65(8):1091–1095.

Madhurananda Pahar, Fuxiang Tao, Bahman Mirheidari, Nathan Pevy, Rebecca Bright, Swapnil Gadgil, Lise Sproson, Dorota Braun, Caitlin Illingworth, Daniel Blackburn, et al. 2025. Cognospeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech. In *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*, pages 1–7. IEEE.

Jungyeul Park and Francis Tyers. 2019. A new annotation scheme for the sejong part-of-speech tagged corpus. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 195–202.

Carrie D Patnode, Leslie A Perdue, Rebecca C Rossom, Megan C Rushkin, Nadia Redmond, Rachel G Thomas, and Jennifer S Lin. 2020. Screening for cognitive impairment in older adults: updated evidence report and systematic review for the us preventive services task force. *JAMA*, 323(8):764–785.

Ronald C. Petersen. 2016. Mild cognitive impairment. *CONTINUUM: Lifelong Learning in Neurology*, 22(2):404–418.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

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*.

Miguel Ramos-Henderson, Carlos Calderón, and Marcos Domic-Siede. 2025. Education bias in typical brief cognitive tests used for the detection of dementia in elderly population with low educational level: a critical review. *Applied Neuropsychology: Adult*, 32(1):253–261.

Sujata Ray and Susan Davidson. 2014. Dementia and cognitive decline. a review of the evidence. *Age UK*, 27:10–12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. 2025. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, 3(2):100205.

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 65(2):519–542.

Vanessa Taler and Natalie A Phillips. 2008. Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556.

Vinh Quang Tran and Haewon Byeon. 2024. Predicting dementia in parkinson's disease on a small tabular dataset using hybrid lightgbm–tabpfn and shap. *Digital Health*, 10:20552076241272585.

Paula T Trzepacz, Helen Hochstetler, Shufang Wang, Brett Walker, Andrew J Saykin, and

Alzheimer's Disease Neuroimaging Initiative. 2015. Relationship between the montreal cognitive assessment and mini-mental state examination for assessment of mild cognitive impairment in older adults. *BMC Geriatrics*, 15(1):107.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott W Linderman. 2021. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pages 4738–4750.

Tian Zheng, Xurong Xie, Xiaolan Peng, Hui Chen, and Feng Tian. 2024. Alzheimer's disease detection based on large language model prompt engineering. In *International Conference on Social Robotics*, pages 207–216. Springer Nature Singapore.

Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth. 2022. Domain-aware intermediate pretraining for dementia detection with limited data. In *Interspeech*, volume 2022, page 2183.

Appendix 1: Prompt templates

All LLM-based classifications use the same system instruction and output constraint, differing only in the data fields provided to the model. The system prompt and output format are shared across all three variants:

System Prompt (all variants)

You are a binary classifier for a research dataset (non-diagnostic). Use only the provided transcript and/or linguistic metrics. Inputs may be English, Slovene, or Korean; treat multilingualism, accent, dialect, and topical content as neutral. Ignore demographic/identity attributes and stereotypes. Assume no class base-rate. Do not reveal reasoning.

Output: Exactly one word — `Control` or `Patient`.

The three prompt variants differ in which data fields are included in the query block:

Variant 1 — Transcript-only

`[DATA INPUT FOR ID: {id} Language: {language}]`

`[TRANSCRIPT]`
{transcript_patient}

`[INSTRUCTIONS]`
Classify strictly as "Control" or "Patient" using only evidence present in the provided fields. Output exactly one word.

`[OUTPUT FORMAT]`
Return exactly one word: Control or Patient.

Variant 2 — Linguistic-only

`[DATA INPUT FOR ID: {id} Language: {language}]`

`[LINGUISTIC METRICS]`
- Speech Rate: *{value}*
- Ttr: *{value}*
- Noun Ratio: *{value}*
- Verb Ratio: *{value}*
- Pronoun Ratio: *{value}*
- Pronoun To Noun Ratio: *{value}*
- Mean Frequency: *{value}*
- Coherence: *{value}*
- Repetitiveness: *{value}*
- Idea Density: *{value}*
- Syntactic Complexity: *{value}*

`[INSTRUCTIONS]`
Classify strictly as "Control" or "Patient" using only evidence present in the provided fields. Output exactly one word.

`[OUTPUT FORMAT]`
Return exactly one word: Control or Patient.

Variant 3 — Full-data (Transcript + Linguistic Metrics)

`[DATA INPUT FOR ID: {id} Language: {language}]`

`[TRANSCRIPT]`
{transcript_patient}

`[LINGUISTIC METRICS]`

- Speech Rate:	*{value}*
- Ttr:	*{value}*
- Noun Ratio:	*{value}*
- Verb Ratio:	*{value}*
- Pronoun Ratio:	*{value}*
- Pronoun To Noun Ratio:	*{value}*
- Mean Frequency:	*{value}*
- Coherence:	*{value}*
- Repetitiveness:	*{value}*
- Idea Density:	*{value}*
- Syntactic Complexity:	*{value}*

`[INSTRUCTIONS]`
Classify strictly as “Control” or “Patient” using only evidence present in the provided fields. Output exactly one word.

`[OUTPUT FORMAT]`
Return exactly one word: Control or Patient.

For few-shot variants, a `[EXAMPLES (labeled)]` block is prepended before the query, containing $2k$ demonstration cases (one per support sample), each formatted identically to the query block above and annotated with their ground-truth label (`Label: Control` or `Label: Patient`).

The Icelandic Language Biobank: Data Collection through a Clinical Analysis Platform

Iris Nowenstein[1], **Naizeth Núñez Macías**[2], **Gunnar Thor Örnólfsson**[1], **Stefán Ólafsson**[2], **Bryndís Bergþórsdóttir**[1], **Iðunn Kristínardóttir**[1], **Hinrik Hafsteinsson**[1]
[1]University of Iceland, [2]Reykjavík University
{irisen, gunnarthor, brynberg, idunnkristinar, hinhaf}@hi.is,
{naizeth23, stefanola}@ru.is

Abstract

Recent work on clinical applications of language technology shows considerable potential for people with speech and language symptoms and disorders, including for the diagnosis and monitoring of diseases and disorders as well as the development of novel communication aids. This has resulted in a variety of digital health tools becoming accessible, including personalized automatic speech recognition for disordered speech and the monitoring of disease progression in neurodegeneration through language samples. Currently, these tools are almost exclusively accessible to speakers of high-resource languages. A major hurdle for small, lower-resourced language communities in this context is the creation of clinical language corpora. We describe ongoing efforts to build the necessary infrastructure for clinical speech and language data collection in Iceland through the Icelandic Language Biobank, a resource that leverages collaboration with clinicians and robust linguistically-informed data collection against data scarcity.

Keywords: clinical language corpora, speech and language disorders, Icelandic

1. Introduction

Advances in language technology over the last decade have led to a significant body of research exploring clinical applications of automatic speech and language analysis. In the context of speech and language symptoms and disorders, two main types of applications rely on clinical corpora.

The first is diagnosis and monitoring, primarily in the context of neurodegenerative diseases such as Alzheimer's, Frontotemporal Dementia, Parkinson's and ALS (e.g., Cho et al., 2024, Shellikeri et al., 2024, Cao et al., 2025), where language sample analysis (based on e.g. picture descriptions) might yield cost-effective, person-centered and non-invasive endpoints for early screening and treatment efficacy assessments, including in drug trials (Robin et al., 2023). Automatic language sample analysis additionally has a long tradition in the context of developmental language disorders and is particularly valuable for the evaluation of bi- and multilingual children (Ortiz et al., 2024), despite challenges in successful technological transfer to clinicians (Klatte et al., 2022, Liu et al., 2023).

Although we focus on conditions which affect speech and language in the current paper, it is important to note that automatic speech and language analysis has also shown potential for diagnosis and monitoring in other clinical contexts (e.g. Malgaroli et al., 2023, Lombardo et al., 2025)

The second application consists of new technology for alternative and augmentative communication (AAC) aids, such as personalized voice synthesis and speech recognition for disordered speech (e.g. MacDonald et al., 2021, Hasegawa-Johnson et al., 2024, Hyppa-Martin et al., 2024). In both diagnosis/monitoring and AAC applications, a range of digital health tools have become available to users, but only for English or a few other high-resource languages. García et al. (2023) point to the ubiquity of English in the field of speech and language markers of neurodegeneration and call for linguistically diverse research, as well as equitable access to novel clinical instruments. It is fairly straightforward to extend this call to action to the field of communication aids based on language technology.

The current paper describes our attempt to answer the call for Icelandic, a low-to-medium resource language (Daðason and Loftsson, 2024) in a small language community of approximately 400,000 speakers. This is done through the creation of the Icelandic Language Biobank (ILB). In contrast with the most comprehensive data collection efforts in high-resource languages (e.g. Hasegawa-Johnson et al., 2024, Kourtis, 2025), the ILB will contain language samples for both AAC and diagnosis/monitoring in order to maximize data exploitation. While this increases the data management challenges, we argue that it is a necessary counterweight to the data scarcity facing the Icelandic language community, which is exacerbated in a clinical context.

Data will primarily be collected through a web-based semi-automatic linguistic analysis platform, ALDA (Automatic Linguistic Data Analysis), designed for and co-created with speech-language pathologists/therapists (SLPs). The purpose of this

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 13–23
12 May 2026.

collaborative process is to ensure successful technological transfer to the clinical context, making the use of our speech and language processing pipeline accessible to clinicians through a user-interface which is tailored to their clinical practice. SLPs are experts in the clinical analysis of speech and language and can therefore benefit greatly from the augmentation of their perceptive and/or manual analysis of language samples through the means of language technology (Klatte et al., 2022, Lian et al., 2025). Similarly, their expertise entails the possibility of direct manual corrections in a clinical context, enhancing analysis quality (hence, the platform is semi-automatic). Creating a platform which combines accessible clinician-centered tools and a data sharing infrastructure could therefore create incentives for durable and sustainable clinician-led data collection, another possible counterweight to data scarcity.

In the current paper, we present the process of building the ILB and SLP-oriented platform, discuss challenges and argue for the crucial role of linguistic knowledge in the endeavor of promoting equitable access to healthcare solutions based on language technology, particularly in the context of language-specific manifestations of disorders and diseases.

Finally, we draw on insights based on corpus linguistics (e.g. Gries, 2010, Wolfer and Koplenig, 2025) and argue that the lack of knowledge on clinically relevant linguistic features for low-to-medium resource languages such as Icelandic can in part be compensated for with the collection of larger and more varied language samples, contra the current trend of decreasing sample length in clinical settings to 1-5 minutes (Petti et al., 2023). Recent ideas about leveraging data from wearables and mobile devices (Kourtis et al., 2019), as has in part been done in the context of language acquisition research (Blom et al., 2023), might therefore be considered particularly beneficial for lower-resource languages. This includes the latest developments of analyzing typing behavior, or "smartphone keyboard input patterns to detect early signs of cognitive impairment" (Samsung Newsroom, 2025).

2. Related work

2.1. State of the art in high-resource contexts

Clinical speech and language corpora, particularly in the context of communication disorders, are not only lacking in less-resourced languages. Even for English, considerable efforts and resources are currently being dedicated to clinical speech and language data collection. We describe two such initiatives below, the Speech Accessibility Project and SpeechDx, as well as the web-based application TELL, which extracts speech and language markers of neurodegeneration and is designed both for clinicians and researchers. We consider these three examples as the current state of the art for clinical speech corpora infrastructure in high-resource languages and use them as a references for the Icelandic Language Biobank and our web-based SLP-oriented analysis platform.

The **Speech Accessibility Project** (Hasegawa-Johnson et al., 2024) is a research initiative funded by Amazon, Apple, Google, Meta, Microsoft and nonprofit organizations. Its aim is to improve automatic speech recognition for non-standard (English) speech by collecting more diverse data through crowdsourcing. Currently, the Speech Accessibility Project collects speech data from paid volunteers (from the U.S., Canada and Puerto Rico) with a variety of speech patterns or disorders and compiles them into an anonymized dataset. Participants can join the project through the web page. As of February 2026, the project had collected at least 1500 hours of recorded speech from more than 1000 participants, including people with Parkinson's disease, Down syndrome, Cerebral Palsy, amyotrophic lateral sclerosis (ALS), and people who have had a stroke. The data contains sentences read out loud that correspond to computer commands, sentences read out loud from (sometimes simplified) novels, and spontaneous speech comprising answers to questions about culture or daily life. Companies and researchers can request access to the corpora created in this project. In addition to the audio files, transcripts, original speech prompts and a subset of the corpora annotated by SLPs are also available. The initiative builds on the experience from Google's Project Euphonia (MacDonald et al., 2021, Tobin and Tomanek, 2022) which laid the groundwork for Project Relate, an Android app with personalized speech recognition for non-standard speech, offering both transcription and resynthesis to aid communication. Importantly, deriving speech markers of diseases and disorders is not one of the aims of the Speech Accessibility Project.

Speech-based biomarkers are on the other hand the main focus of **SpeechDx** (Kourtis, 2025), a project in which the goal is to create a longitudinal dataset (spanning three years) for the diagnosis of Alzheimer's disease and related dementias through speech samples. The goal is to cover English, Spanish and Catalan and to link comprehensive clinical information to the participants' speech data. The dataset includes samples obtained through different elicitation tasks, such as picture descriptions, open-ended questions, story recall and storytelling. The data will be hosted by the Alzheimer's Disease Data Initiative and will be made available to researchers approved by a committee in a pro-

tected, controlled environment.

The **TELL** application (García et al., 2024b) provides "robust speech biomarkers for clinical and research purpose" and has mostly been deployed in the context of neurodegeneration. The first deployment was available for English, Spanish, French and Portuguese. Although its latest version also enables data collection for German, Italian, Quechua, Kiswahili and Tagalog (García et al., 2024a), language-specific features (such as POS tags, semantic granularity etc.) are only (automatically) extracted for the higher-resource languages. The platform is designed to be used by clinicians as well as researchers, but SLPs are not the main target group. This is comparable to Open Brain AI (Themistocleous, 2024), a relatively new computational platform which currently provides tools for automatic language sample analysis in 15 languages and is also designed for both clinicians and researchers.

To the best of our knowledge, no comprehensive automatic speech and language analysis platform is designed for the needs of SLPs both in the fields of developmental and acquired communication disorders. One of the key challenges is the technical skills needed to implement available tools in clinical practice. For instance, the Batchalign pipeline was developed for the automatic transcription and analysis of clinical samples in the CHAT (Codes for the Human Analysis of Talk) format using the software program CLAN (Computerized Language Analysis, Liu et al., 2023). One of the main goals of Batchalign was to reduce the time needed to transcribe raw audio files by enabling clinicians to generate an automatic transcription, which only had to be manually corrected. However, another study found that SLPs did not experience a reduction in the time needed to perform language sample analysis despite receiving tailored training (Klatte et al., 2022). In fact, the participants reported a lack of knowledge and skills as a barrier to using tools such as Batchalign. These results highlight the need for user-friendly software and the incorporation of SLPs in the tool design process. Additionally, we believe a crucial component, particularly for lower-resource languages with fewer tools and higher error rates, is to enable clinicians such as SLPs to correct the automatic analysis.

2.2. The Icelandic landscape

Not unexpectedly, Icelandic implementations within state-of-the-art tools and datasets for digital health are limited. For example, there are no direct analogs to the aforementioned TELL, Open Brain AI and CHAT in Iceland. This is in part due to the lack of datasets, research and development in the domain of clinical language technology for Icelandic.

As is the case in many small(er) language communities, collecting clinical linguistic data for Icelandic presents a number of challenges. This is not only due to the limited number of speakers, but also to the lack of infrastructure to safely collect, store, and share language samples, as well as the lack of focused research on clinical populations speaking those languages. In this sense, there is a real risk that Icelandic will fall further behind with regards to the development and accessibility of automatic speech and language analysis tools and datasets for digital health.

As briefly mentioned, an inherent difficulty for small language communities like Icelandic is the overall low number of individuals diagnosed with the targeted diseases and disorders. For example, there are approximately 20-30 people with ALS in Iceland at any given time (MND Iceland). These individuals face the same difficulties as people with ALS in larger communities, but the development of novel, language-specific solutions is limited by the fact that data collection can only be obtained through a very low number of individuals. We believe this entails that any novel data for Icelandic needs to be utilized to the fullest.

Fortunately, there are already projects that have gathered Icelandic language samples for uses in clinical language technology development. Specifically, these projects gathered language samples from people with Mild Cognitive Impairment and mild dementia due to Alzheimer's disease (Curcic et al., 2022, Callegari et al., 2023 and Nowenstein et al., 2024). However, there are no accessible corpora with language samples from other clinical groups where automatic language sample analysis has proven useful, e.g., people with Parkinson's, ALS, Frontotemporal Dementia and aphasia and/or motor speech disorders following a stroke, as well as language samples from children with developmental language disorders.

The first results derived from the two Alzheimer's disease datasets described above highlight the importance of taking into account the characteristics of individual languages when generalizing previous results and extending the use of clinical language technology to new contexts. For example, Callegari et al. (2024) show that the frequency of various features will vary greatly across discourse contexts in different types of language samples, both within features commonly extracted in previous research (e.g. verb rate) and features more specific to the Icelandic language (e.g. subjunctive rate).

Another recent project comparing speech and language markers of neurodegeneration across English, Korean and Icelandic found that language-specific features, such as the use of case marking in Icelandic, can differentiate between clinical groups and healthy controls (Nowenstein et al.,

2025). Similar results have been found in the field of Developmental Language Disorders (DLD), where research on Icelandic indicates that the influential view (Leonard, 2014) of morphological errors as a hallmark of DLD does not necessarily hold for languages with richer morphological systems (Thordardottir, 2016).

3. Building the Icelandic Language Biobank

The Icelandic Language Biobank (ILB) is a three year initiative funded by The Strategic Research and Development Programme for Language Technology within the Icelandic Research Fund. Its preparation was also funded by the Language Technology Programme for Icelandic. The ILB is an attempt to answer the call for increased linguistic diversity in clinical language technology (García et al., 2023) and could serve as a proof of concept for other small language communities.

Our goal is to collect clinical language samples from speakers of Icelandic (children and adults, mono- and multilingual) in order to improve access to healthcare solutions based on language technology. We focus on the manifestations of developmental language disorders and neurodegeneration in Icelandic and build the infrastructure for the collection and preservation of clinical speech and language corpora for Icelandic in a broad sense. A further goal of the ILB infrastructure is to allow for the collection, annotation, and analysis of samples in other languages.

We collaborate with the leadership of the Speech Accessibility project (Hasegawa-Johnson et al., 2024) and the DELAD initiative within CLARIN. DELAD facilitates the sharing of corpora of speech of individuals with communication disorders among researchers in a GDPR compliant way and at secure repositories in the CLARIN infrastructure (Lee et al., 2024).

The design of the ILB is motivated by the small size and lack of resources of the Icelandic language community. We try to combine the approaches of initiatives such as the Speech Accessibility Project, SpeechDx and TELL into one centralized solution to maximize data exploitation. This means that the same speaker can provide data for improved communication aids (e.g. ASR for disordered speech) and diagnosis/monitoring (e.g. digital biomarkers of neurodegenration) through a clinician (SLP) who has access to automatic speech and language data analysis with our data collection and analysis platform. Additionally, SLPs' expertise in speech and language means they are particularly powerful collaborators, including in the context of transcription and annotation correction.

One of the key motivations behind the ILB is the current lack of knowledge when it comes to the direct clinical application of language technology, particularly when it comes to the interpretability of the information retrieved through automatic speech and language analysis. Even though automatic analysis tools may reduce the workload of SLPs (Liu et al., 2023), it is not always clear how SLPs can interpret these new measures within evidence-based practice (Lindsay et al., 2021, Yeung et al., 2021).

Another area where knowledge is lacking is the transfer of research findings from English to other languages, as it is known that the manifestations of language disorders depend to some extent on the specific characteristics of different languages. For instance, research has found that there is an increase in the rate of pronouns in English in Alzheimer's disease (see Petti et al., 2020, Robin et al., 2021 and Cho et al., 2022), while the reverse pattern was found for pro-drop languages such as Bengali (Bose et al., 2021). An overemphasis on English can therefore produce a biased view of what characterizes language symptoms and disorders in general (Thordardottir, 2016, García et al., 2023), which underlines the need for typologically diverse clinical linguistic data and the necessary infrastructure to collect it. This is relevant, among other things, for improved diagnosis of developmental disorders in multilingual children.

3.1. Infrastructure and data management

Data management is one of the key challenges when building infrastructure for the collection and preservation of clinical speech and language data. The ILB will draw from the various data collection initiatives described in section 2.1, complying with e.g. the GDPR and taking into consideration other developments at the EU level, such as the AI and Data Acts as well as the European Health Data Space Regulation. The project also needs to comply with Icelandic legislature on biobanks and clinical datasets and request approval from The National Bioethics Committee of Iceland. The project is hosted at the University of Iceland and has benefitted from consultations with the university's IT departments as well as legal counsel specialized in personal data protection.

As is detailed in García et al. 2024a, the TELL application is deployed on Amazon Web Services (AWS) with data handling through Amazon's RDS PostgreSQL database, audio files stored in AWS S3 and patient health information encrypted on AWS Key Management Service. Our current data management plan is similar in structure but is currently hosted completely locally at the University of Iceland given Icelandic legislature which dictates that clinical data should in general be exclusively processed and stored domestically. This will poten-

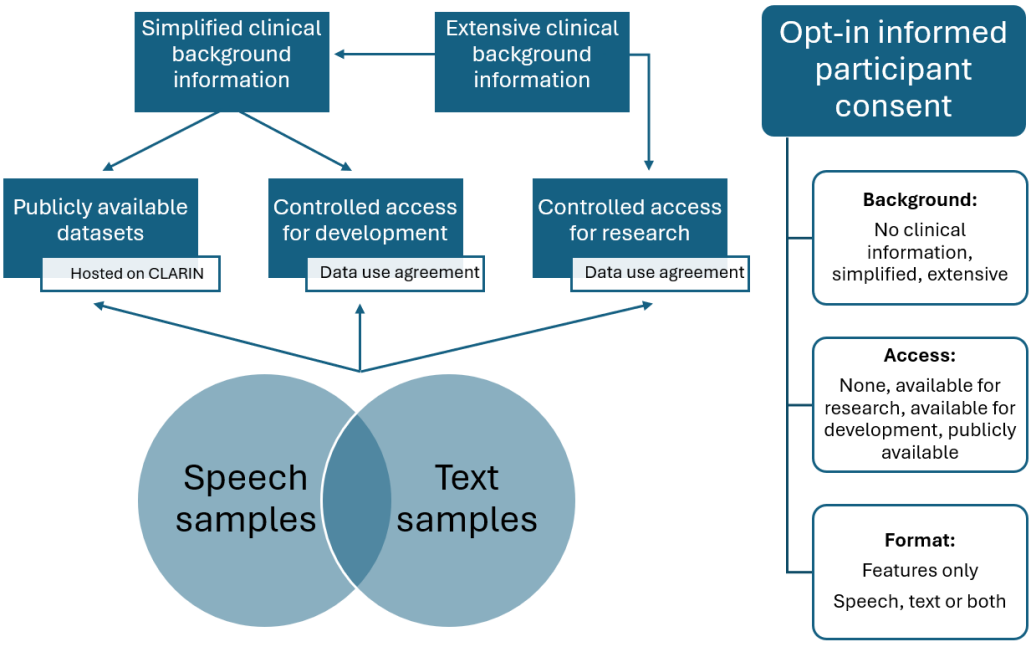


Figure 1: Icelandic Language Biobank consent and data sharing infrastructure.

tially cause scalability challenges with increased data collection efforts, a complication we are currently addressing.

When data has been collected through the web platform, we transfer it to a Nextcloud server hosted within the IREI cluster at the University of Iceland through encrypted data transfer. At the University of Iceland, the data is reviewed and personally identifying information (PII) is manually removed. Participants' identification numbers are stored separately from the deidentified, pseudonymized data and not shared unless a new approval by The National Bioethics Committee has been granted. As shown in Figure 1, the ILB will adopt a layered approach with opt-in (and opt-out) informed participant consent for the deidentified data, meaning that the project will share data in a variety of forms.

This means that participants can opt into and out of different levels of data sharing for different groups, allowing for example developers to have controlled access to simplified or no clinical information through a data use agreement while researchers can access more extensive clinical information, also under a data use agreement. Furthermore, participants will have control over what information from their sample is shared, ranging from audio recordings to feature values only. This entails providing participants with comprehensive information, for example regarding their voice potentially being identifiable to their acquaintances.

Since the ILB aims to collect data from a number of vulnerable groups and/or protected entities, precautions are in order to ensure that the consent can truly be considered informed. In some cases, this will be in the form of parental or guardian informed consent as well as participant assent with a simplified information form. To ensure appropriate informed consent, the ILB builds on the DELAD resources presented in Lee et al. (2024) as well as the PEEC (Protected Entities Ethics Checklist) for collecting speech data from vulnerable clinical populations (Choi et al., forthcoming). For instance, according to the PEEC framework, children's consent should be tailored to their developmental stage. Children above 7 may provide written assent, while the affirmative willingness of younger participants may be provided after a simplified, verbal explanation (Choi et al., forthcoming). At any given point, participants may withdraw their consent and demand the deletion of their data.

Data sharing will be handled under restricted licenses within the DELAD-initiative through CLARIN, as DELAD is linked to CLARIN's Knowledge Centre for Atypical Communication Expertise (ACE) for making corpora of speech of individuals with communication disorders (CSD) available through The Language Archive (TLA) at the Max Planck Institute in Nijmegen (a CLARIN Data Centre) and CMU's Talkbank (Clinical Banks). We therefore provide an infrastructure which makes it possible to ensure the data of the ILB will be as accessible as possible while guaranteeing participants' data privacy according to their wishes.

3.2. ALDA: web-based semi-automatic linguistic analysis platform

A key driver of sustainable and longitudinal collection of clinical data within the ILB is collaboration with SLPs and other clinicians. For such collaboration to be gainful for all parties, we strive to provide utility to the clinicians in exchange for their contribution to data collection. Therefore, the development team for the ALDA platform includes two SLPs, one working with pediatric populations and the other specializing in neurodegeneration.

ALDA (Automatic Linguistic Data Analysis) is a web-based semi-automatic linguistic analysis platform that allows clinicians specialized in speech and language disorders to perform Language Technology-aided analysis of clinical language samples through a user-friendly interface. The platform also allows the clinicians to manage their clients' language samples in a centralized storage space and track indicators of diseases and disorders over time. The platform is currently being developed with funding from the Language Technology Programme for Icelandic.

As illustrated in Figure 2, the SLP records a speech sample using ALDA, for example a picture description or story recall, or uploads a previous recording. The sample is then transcribed and diarized using automatic speech recognition and corrected manually as necessary. This step of language sample analysis has so far been performed fully manually by Icelandic SLPs. ALDA also supports tasks for children and adults with feature bundles for different types of speech and language disorders.

ALDA utilizes various speech and language processing tools to perform automated analyses of the speech sample, e.g. for ASR and speaker diarization, POS-tagging and parsing. The current version of the platform integrates WhisperX (Bain et al., 2023) for VAD (Voice Activity Detection), ASR, forced alignment and speaker diarization. For optimal results, versions of Whisper and Wav2Vec2 which have been fine-tuned on Icelandic speech corpora are used (Radford et al., 2023, Baevski et al., 2020, Mena et al., 2024). A PoS tagger using a fine-grained morphological tagset (Jónsson et al., 2021) is used to extract morphosyntactic information from the transcribed speech. We stress that the pipeline is under development and needs further testing in terms of e.g. preprocessing steps with diverse data sources as well as error rates for children's voices and disordered speech. Our text processing pipeline will also keep evolving. Although preliminary findings suggest acceptable POS-tagging accuracy in clinical conversational language samples, a lot of challenges remain for syntactic parsing (not unexpectedly, see Agmon et al., 2026 for English). Finally, the current pipeline only supports the analysis of Icelandic language samples.

After data collection, recordings and analysis results can then be saved in the SLPs' secure storage space within the platform. ALDA also has a built-in data collection functionality that enables SLPs using the platform to invite clients to receive information about the ILB project. When a client is interested in participating and informed consent has been obtained through an electronic signature on the web page connected to the Icelandic Language Biobank, the existing participant's language samples can be transferred to the ILB. This data collection sets the ILB apart methodologically, as the data collection tool itself is an aid in the current workflow of SLPs who already record language samples within their clinical practice. With ALDA, SLPs have a tool for immediate use in practice, not only after the data have been processed for further research and technology transfer.

In order to involve further SLPs into the development process, two focus groups with Icelandic SLPs have been conducted, confirming the group's interest in using language technology to facilitate language sample analysis. The participants also expressed enthusiasm for the use of language technology for communication aids and were positive towards collaboration with researchers, but emphasized the importance of receiving detailed instructions and a clear and accessible protocol. Interestingly, SLPs did not show enthusiasm for language sample analysis in languages they do not speak and cited the importance of being able to verify the analysis to interpret the results.

In addition to the focus groups, 35 practicing SLPs participated in a survey we conducted about language sample analysis. 91.4% of the participants considered language samples as a useful tool, the rest considered it useful in certain cases. Additionally, 91% of our participants said they would use language samples more frequently if processing them took less time and they had access to better tools for it. Currently, user testing for the platform is ongoing. Until the end of 2026, we aim to identify existing barriers, technical difficulties and SLPs' concerns in order to iteratively improve the platform before starting data collection. Although the current pipeline only supports Icelandic, the platform can of course be used to record and store samples in other languages.

3.3. Data collection within the ILB

In the second year of the Icelandic Language Biobank project (January 2027), its infrastructure will be ready and an initial data collection and analysis phase will begin. Within the project, data collection will be conducted through clinician-led participant recruitment as well as crowdsourcing efforts.

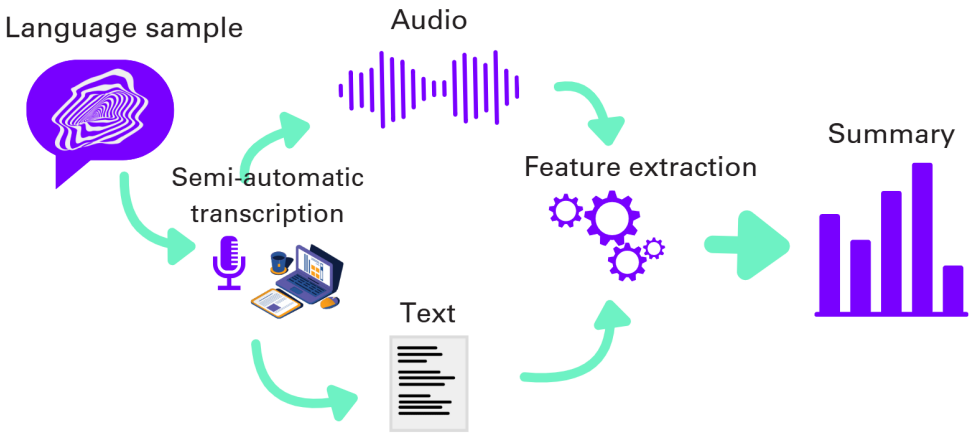


Figure 2: High-level diagram of the flow of language samples through the analysis pipeline. Features are extracted from both the diarized, text-aligned recording and the text transcript.

Both the target dataset size and characteristics of participants will be established during the first three months of data collection based on funding availability and clinician adoption of the web platform, but the primary goal of the project is to build the necessary infrastructure for continuous data collection.

The crowdsourcing will target clinical groups (as has been done in the Speech Accessibility Project) but mainly people without communication disorders. This is because having robust control group data is a crucial step for the use of automatic speech and language analysis in a clinical setting, where situating individual patients within well-established norms is an important component of diagnosis. Indeed, 88% of participants in the aforementioned SLP survey said they would use language samples more if they had better norms as a comparison.

The clinician-led recruitment of participants from clinical groups will be based on the use of the ALDA platform, with the ILB project providing training for SLPs in the form of in-person courses and online instructional content. The first phase of data collection within the ILB project will be centered around developmental language disorders in mono- and multilingual children as well as speech and language disorders in neurodegeneration, using classic language sampling methods such as picture descriptions and story recall.

Although the initial data collection is focused on clinician-led recruitment and crowdsourcing, the goal of the ILB project still is to establish a durable infrastructure for any type of clinical speech and language data, including new types of language samples collected through e.g. wearables and mobile devices. Considering common knowledge about the positive effects of increasing corpus size to obtain more representative samples of language use (e.g. Gries, 2010), we believe that efforts should be made to increase clinical language sample length without increasing clinician burden. Currently, the direction in the literature is to decrease sample length (Petti et al., 2023), but larger clinical language samples from individuals might be an important way of countering the data scarcity inherent to smaller language communities. We also believe there might be benefits to expanding clinical language sampling to analyses of participants' written language output and the ILB will therefore also accommodate the storage of written language samples. Scaling from relatively homogeneous, short spoken language samples to more diverse and extensive types of data will present problems we have not solved yet but aim to address, both in terms of storing and processing/analyzing the data. Nevertheless, we believe academic research in small language communities should strive to accommodate as much data diversity as possible.

4. Conclusion

Recent applications of language technology show that children and adults with communication disorders and the clinicians who treat them stand to benefit considerably from successful technological transfer of speech and language processing tools. A necessary step in that process is the collection of speech and language samples from people with communication disorders. In the context of less-resourced languages, particularly in small language communities where data scarcity will be problematic, we suggest building data infrastructure which will make it possible to collect data both for (1) diagnosis/monitoring of diseases and disorders (including clinical information about the participants) and (2) communication aids, including better

speech recognition for disordered speech.

We presented our approach for this kind of infrastructure in Iceland through the creation of the Icelandic Language Biobank, a project which also includes comprehensive collaboration with clinicians by providing them with tools for data analysis on a platform which additionally serves as a data collection point. We believe that this combination of a comprehensive one-stop approach to corpora of speech of individuals with communication disorders and bilateral collaboration with clinicians will provide some of the necessary counterweight to data scarcity in small less-resourced language communities.

Another way to approach the problem is through longer clinical language samples, possibly leveraging wearables and mobile devices and going beyond speech samples to include individuals' written outputs as well. For this kind of research on large clinical language samples, it is even more important to build data infrastructure where data security and personal privacy are guaranteed.

5. Acknowledgements

The projects presented in the current paper were funded by The Strategic Research and Development Programme for Language Technology within the Icelandic Research Fund and the Language Technology Programme for Icelandic. We would also like to thank the anonymous reviewers for their valuable feedback and comments.

6. Ethical considerations

Building the Icelandic Language Biobank relies on an in-depth mapping of ethical consideration. This is why we base our work on the Protected Entities Ethics Checklist (PEEC, Choi et al., forthcoming), a comprehensive framework specifically designed for researchers collecting speech and language data from clinically vulnerable populations, including children, elderly adults with cognitive changes, individuals with communication disorders, and marginalized communities.

Finally, although the motivation for the Icelandic Language Biobank is centered around stakeholder needs (ensuring accessibility to clinical language technology regardless of the language people speak), it still is crucial to consult with stakeholders, particularly in vulnerable clinical populations, to ensure their perspectives and interests are embedded into the project they contribute to.

7. Bibliographical References

Galit Agmon, Sunghye Cho, Sharon Ash, Katheryn A. Q. Cousins, Kaj Blennow, Henrik Zetterberg, Leslie M. Shaw, Sameer Pradhan, Yoon Duk Kim, Mark Y. Liberman, David J. Irwin, and Naomi Nevler. 2026. Automatic quantification of syntactic complexity in natural spontaneous speech of people with primary progressive aphasia. *Aphasiology*, 40(3):561–582.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

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.

Elma Blom, Paula Fikkert, Annette Scheper, Merel van Witteloostuijn, and Petra van Alphen. 2023. The Language Environment at Home of Children With (a Suspicion of) a Developmental Language Disorder and Relations With Standardized Language Measures. *Journal of Speech, Language, and Hearing research*, 66(8):2821–2830.

Arpita Bose, Niladri S. Dash, Samrah Ahmed, Manaswita Dutta, Aparna Dutt, Ranita Nandi, Yesi Cheng, and Tina M. D. Mello. 2021. Connected Speech Characteristics of Bengali Speakers With Alzheimer's Disease: Evidence for Language-Specific Diagnostic Markers. *Frontiers Aging Neuroscience*, 13:707628.

Elena Callegari, Iris Edda Nowenstein, Ingunn Jóhanna Kristjánsdóttir, and Anton Karl Ingason. 2024. Automatic Extraction of Language-Specific Biomarkers of Healthy Aging in Icelandic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1915–1924, Torino, Italia. ELRA and ICCL.

Elena Callegari, Agnes Sólmundsdóttir, and Anton Karl Ingason. 2023. The ACoDe Project: Creating a Dementia Corpus for Icelandic. In *Proceedings of CLARIN Annual Conference 2023*, pages 100–105.

Fangyuan Cao, Adam P. Vogel, Puya Gharahkhani, and Miguel E. Renteria. 2025. Speech and language biomarkers for parkinson's disease prediction, early diagnosis and progression. *npj Parkinson's Disease*, 11(1):57.

Sunghye Cho, Katheryn Alexandra Quilico Cousins, Sanjana Shellikeri, Sharon Ash, David John Irwin, Mark Yoffe Liberman, Murray Grossman, and Naomi Nevler. 2022. Lexical and Acoustic Speech Features Relating to Alzheimer Disease Pathology. *Neurology*, 99(4):e313–e322.

Sunghye Cho, Christopher A. Olm, Sharon Ash, Sanjana Shellikeri, Galit Agmon, Katheryn A. Q. Cousins, David J. Irwin, Murray Grossman, Mark Liberman, and Naomi Nevler. 2024. Automatic classification of AD pathology in FTD phenotypes using natural speech. *Alzheimer's & Dementia*, 20(5):3416–3428.

Anna Seo Gyeong Choi, Sunghye Cho, and Iris Nowenstein. PEEC: The Protected Entities Ethics Checklist for Collecting Speech Data from Vulnerable Clinical Populations. Accepted manuscript, *Journal of Speech, Language, and Hearing Research*.

Claire Cordella, Manuel J. Marte, Hantian Liu, and Swathi Kiran. 2025. An Introduction to Machine Learning for Speech-Language Pathologists: Concepts, Terminology, and Emerging Applications. *Perspectives of the ASHA Special Interest Groups*, 10(2):432–450.

Jelena Curcic, Vanessa Vallejo, Jennifer Sorinas, Oleksandr Sverdlov, Jens Praestgaard, Mateusz Piksa, Mark Deurinck, Gul Erdemli, Maximilian Bügler, Ioannis Tarnanas, Nick Taptiklis, Francesca Cormack, Rebekka Anker, Fabien Massé, William Souillard-Mandar, Nathan Intrator, Lior Molcho, Erica Madero, Nicholas Bott, Mieko Chambers, Josef Tamory, Matias Shulz, Gerardo Fernandez, William Simpson, Jessica Robin, Jón G. Snædal, Jang-Ho Cha, and Kristin Hannesdottir. 2022. Description of the Method for Evaluating Digital Endpoints in Alzheimer Disease Study: Protocol for an Exploratory, Cross-sectional Study. *JMIR Research Protocols*, 11(8):e35442.

Jón Daðason and Hrafn Loftsson. 2024. Text filtering classifiers for medium-resource languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15789–15801, Torino, Italia. ELRA and ICCL.

Adolfo M. García, Jessica de Leon, Boon Lead Tee, Damián E. Blasi, and Maria Luisa Gorno-Tempini. 2023. Speech and language markers of neurodegeneration: a call for global equity. *Brain*, 146(12):4870–4879.

Adolfo M. García, Franco J. Ferrante, Gonzalo Pérez, Joaquín Ponferrada, Alejandro Sosa Welford, Nicolás Pelella, Matías Caccia, Laouen Mayal Louan Belloli, Cecilia Calcaterra, Catalina González Santibáñez, Raúl Echegoyen, Mariano Javier Cerrutti, Fernando Johann, Eugenia Hesse, and Facundo Carrillo. 2024a. Toolkit to Examine Lifelike Language v.2.0: Optimizing Speech Biomarkers of Neurodegeneration. *Dementia and Geriatric Cognitive Disorders*, 54(2):96–108.

Adolfo M. García, Fernando Johann, Raúl Echegoyen, Cecilia Calcaterra, Pablo Riera, Laouen Belloli, and Facundo Carrillo. 2024b. Toolkit to Examine Lifelike Language (TELL): An app to capture speech and language markers of neurodegeneration. *Behavior Research Methods*, 56(4):2886–2900.

Stefan Th. Gries. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez and Moisés Almela, editors, *A mosaic of corpus linguistics: selected approaches*, pages 269–291. Peter Lang, Frankfurt.

Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwilling, Marie Moore Channell, Laura Mattie, Heather Hodges, Lorraine Ramig, Mary Bellard, Mike Shebanek, Leda Sarı, Kaustubh Kalgaonkar, David Frerichs, Jeffrey P. Bigham, Leah Findlater, Colin Lea, Sarah Herrlinger, Peter Korn, Shadi Abou-Zahra, Rus Heywood, Katrin Tomanek, and Bob MacDonald. 2024. Community-Supported Shared Infrastructure in Support of Speech Accessibility. *Journal of Speech, Language, and Hearing Research*, 67(11):4162–4175.

Jolene Hyppa-Martin, Jason Lilley, Mo Chen, Jaclyn Friese, Corinne Schmidt, and H Timothy Bunnell. 2024. A large-scale comparison of two voice synthesis techniques on intelligibility, naturalness, preferences, and attitudes toward voices banked by individuals with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication*, 40(1):31–45.

Inge S. Klatte, Vera Van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language Sample Analysis in Clinical Practice: Speech-Language Pathologists' Barriers, Facilitators, and Needs. *Language, Speech, and Hearing Services in Schools*, 53(1):1–16.

Lampros C. Kourtis. 2025. Speechdx: A gold-standard speech-and-language dataset for prognostic ad biomarker development. *Alzheimer's & Dementia*, 21:e104638.

Lampros C. Kourtis, Oliver B. Regele, Justin M. Wright, and Graham B. Jones. 2019. Digital

biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *npj Digital Medicine*, 2(1):9.

Alice Lee, Nicola Bessell, Henk Van Den Heuvel, Katarzyna Klessa, and Satu Saalasti. 2024. The DELAD initiative for sharing language resources on speech disorders. *Language Resources and Evaluation*, 58(3):865–879.

Laurence B. Leonard. 2014. Children with specific language impairment and their contribution to the study of language development. *Journal of Child Language*, 41(S1):38–47.

Jiachen Lian, Xuanru Zhou, Chenxu Guo, Zongli Ye, Zoe Ezzes, Jet M.J. Vonk, Brittany Morin, David Baquirin, Zachary Miller, Maria Luisa Gorno-Tempini, and Gopala Krishna Anumanchipalli. 2025. Automatic Detection of Articulatory-Based Disfluencies in Primary Progressive Aphasia. *IEEE Journal of Selected Topics in Signal Processing*, 19(5):810–826.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021. Language Impairment in Alzheimer's Disease-Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, 13:642033.

Houjun Liu, Brian MacWhinney, Davida Fromm, and Alyssa Lanzi. 2023. Automation of Language Sample Analysis. *Journal of Speech, Language, and Hearing Research*, 66(7):2421–2433.

Clara Lombardo, Giulia Esposito, Silvia Carbone, Salvatore Serrano, and Carmela Mento. 2025. Speech analysis and speech emotion recognition in mental disease: a scoping review. *Frontiers in Psychology*, 16.

Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, Jordan R. Green, and Katrin Tomanek. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. In *Interspeech 2021*, pages 4833–4837. ISCA.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Carlos Mena, Þorsteinn Daði Gunnarsson, and Jon Gudnason. 2024. SamróMur MilljóN: An ASR corpus of one million verified read prompts in Icelandic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14305–14312, Torino, Italia. ELRA and ICCL.

MND Iceland. Hvað er MND? [What is ALS?]. Accessed October 24, 2025.

Iris Nowenstein, Min Seok Baek, Bryndís Bergþórsdóttir, Daria Birju, Elena Callegari, Hinrik Hafsteinsson, Anton Karl Ingason, María K. Jónsdóttir, Ashley Keaton, Sungoo Kim, Seohee Kim, Louis Kwak, Judith Neugroschl, Caitlin Richter, Mary Sano, Truda Silberstein, Jón Snædal, Laila Soleimani, Gunnar Thor Örnólfsson, Carolyn Zhu, and Sunghye Cho. 2025. Speech and language markers of cognitive decline and neurodegeneration: Generalizability across languages. In *Society for the Neurobiology of Language 17th Annual Meeting*, Gallaudet University.

Iris Nowenstein, Marija Stanojevic, Gunnar Örnólfsson, María Kristín Jónsdóttir, Bill Simpson, Jennifer Sorinas Nerin, Bryndís Bergþórsdóttir, Kristín Hannesdóttir, Jekaterina Novikova, and Jelena Curcic. 2024. Speech and Language Biomarkers of Neurodegenerative Conditions: Developing Cross-Linguistically Valid Tools for Automatic Analysis. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 26–33, Torino, Italia. ELRA and ICCL.

José A. Ortiz, Jessica M. Nolasco, Yi Ting Huang, and Jason C. Chow. 2024. The Use of Language Sample Analysis to Differentiate Developmental Language Disorder From Typical Language in Bilingual Children: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, 67(10):3803–3825.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Ulla Petti, Simon Baker, Anna Korhonen, and Jessica Robin. 2023. How Much Speech Data Is Needed for Tracking Language Change in Alzheimer's Disease? A Comparison of Random Length, 5-Min, and 1-Min Spontaneous Speech Samples. *Digital Biomarkers*, 7(1):157–166.

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*, ICML'23. JMLR.org.

Jessica Robin, Mengdan Xu, Aparna Balagopalan, Jekaterina Novikova, Laura Kahn, Abdi Oday, Mohsen Hejrati, Somaye Hashemifar, Mohammadreza Negahdar, William Simpson, and Edmond Teng. 2023. Automated detection of progressive speech changes in early alzheimer's disease. *Alzheimer's & Dementia*, 15(2):e12445.

Jessica Robin, Mengdan Xu, Liam D. Kaufman, and William Simpson. 2021. Using Digital Speech Assessments to Detect Early Signs of Cognitive Impairment. *Frontiers in Digital Health*, 3:749758.

Samsung Newsroom. 2025. [World Alzheimer's Day] Samsung Research Advances Early Detection of Alzheimer's With Everyday Digital Data.

Sanjana Shellikeri, Sunghye Cho, Sharon Ash, Carmen Gonzalez-Recober, Katheryn A Q Cousins, Corey T McMillan, Lauren Elman, Colin Quinn, Defne A Amado, Michael Baer, David J Irwin, Lauren Massimo, Mark Y Liberman, and Naomi Nevler. 2024. Digital speech markers of cognitive impairment in ALS-FTD spectrum disorders. *Alzheimer's & Dementia*, 20(S2):e089943.

Charalambos Themistocleous. 2024. Open Brain AI and language assessment. *Frontiers in Human Neuroscience*, 18:1421435.

Elin Thordardottir. 2016. Grammatical morphology is not a sensitive marker of language impairment in Icelandic in children aged 4–14 years. *Journal of Communication Disorders*, 62:82–100.

Jimmy Tobin and Katrin Tomanek. 2022. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6637–6641.

Sascha Wolfer and Alexander Koplenig. 2025. Does corpus size influence normalised frequencies? *Corpus Linguistics and Linguistic Theory*.

Anthony Yeung, Andrea Iaboni, Elizabeth Rochon, Monica Lavoie, Calvin Santiago, Maria Yancheva, Jekaterina Novikova, Mengdan Xu, Jessica Robin, Liam D. Kaufman, and Fariya Mostafa. 2021. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimer's Research & Therapy*, 13(1):109.

8. Language Resource References

Jónsson, Haukur Páll and Loftson, Hrafn and Steingrímsson, Steinþór. 2021. *ABLTagger (PoS) - 3.0.0*. Reykjavík University. PID http://hdl.handle.net/20.500.12537/115. CLARIN-IS.

Disfluencies and ASR Performance on Swedish Spontaneous Speech from the 'Trip to Stockholm' Discourse Narrative Task

Dimitrios Kokkinakis, Herbert Lange, Ricardo Muñoz Sánchez
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg, Sweden
{dimitrios.kokkinakis, herbert.lange, ricardo.munoz.sanchez}@svenska.gu.se

Abstract

Automatic Speech Recognition (ASR) offers a scalable and cost-efficient alternative to manual transcription and is becoming increasingly relevant in clinical contexts, particularly for the detection of cognitive decline and mental health assessment. However, current ASR-systems still struggle with spontaneous speech, particularly when processing disfluencies, pauses, and speaker variability that often carry diagnostic value. This study evaluates state-of-the-art open ASR models targeting Swedish using recordings from the "Trip to Stockholm" discourse narrative task which elicits ecologically valid, cognitively demanding speech. Recognition quality is assessed using various metrics, alongside an analysis of linguistic and technical sources of error focused on disfluencies. Our findings show that disfluency-related phenomena degrade recognition performance. Possible post-processing strategies can improve specific error patterns emerging for filled pauses, word repetitions, and self-corrections. The results illustrate both the advances *and* ongoing limitations of ASR for spontaneous Swedish speech, emphasizing the need for models explicitly trained, or fine-tuned, on disfluent data to ensure robustness in clinical and research applications.

Keywords: Whisper, KB-Whisper, speech-to-text, disfluencies, discourse task, cognitive impairment, Swedish

1. Introduction

Automatic Speech Recognition (ASR) is a scalable and cost-efficient technology that addresses the increasing volume of digital content and the corresponding demand for accessible communication. ASR offers several advantages, including low deployment costs, minimal logistical and availability constraints, and the ability to operate continuously without interruption, making it particularly attractive for clinical applications such as mental health assessment and the detection of cognitive decline. An important advantage of ASR is its potential to mitigate the cost and time demands of manual transcription while reducing transcription errors and forms of human-induced bias, such as subjective interpretation and inconsistencies across transcribers. Given the privacy and recording constraints frequently encountered in clinical settings, ASR provides a scalable means of capturing spoken responses for subsequent analysis.

However, ASR errors—particularly failures to accurately capture disfluencies and pauses that are central to cognitive assessment—constitute a persistent and consequential limitation, potentially obscuring dementia-specific speech markers (cf. Li et al., 2024). In this study, we evaluate the performance of automatic speech recognition systems on Swedish speech data, with a specific focus on disfluencies—excluding pauses—by assessing recognition outcomes using Word Error Rate (WER) and complementary metrics. We examine linguistic factors and technical constraints underlying transcription errors and systematically analyze common error patterns in spontaneous speech across different hyperparameter settings. Since performance metrics quantify the deviation of a transcript from a reference, a central aim of this study is to also discuss ways to improve transcription output.

We apply the performance metrics to a realistic dataset by first applying state-of-the-art open ASR models and assessing their ability to handle diverse linguistic structures and speech variability across varying speaker groups. Recordings from the "Trip to Stockholm" task were used as input for evaluating ASR models. This is a spoken discourse task which was modeled after the "Trip to New York" task described in Harris et al. (2008). The spontaneous and variable nature of these speech samples provides a valuable benchmark for evaluating ASR performance in ecologically valid, cognitively demanding, Swedish discourse settings.

2. Background and Related Work

Speech disfluencies — such as repetitions, self-corrections, filled pauses and other interruptions in the flow of speech — are a natural feature of spontaneous language and have been widely studied for their informational value. Additionally, disfluencies are well-established indicators of cognitive and mental health status and have been widely associated with neurodegenerative conditions such as dementia, particularly Alzheimer's disease (AD).

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 24–33
12 May 2026.

Such non-fluent speech patterns have been shown to increase with disruptions in language planning, executive functioning, and cognitive load (Jiang and An, 2025; Clark and Fox Tree, 2002), highlighting how such phenomena relate to cognitive and lexical-semantic impairment in AD (Pistono et al., 2024).

State-of-the-art ASR systems do not adequately capture and label word- and phrase-level disfluencies (Shahla et al., 2022; Cumbal et al., 2024). In addition, in a diachronic context, age correlates with changes in speech rate and lexical complexity, which, in turn, contribute to increased production of disfluency over time, establishing disfluency as a robust marker of age-related linguistic change (Beier et al., 2023). As Nasreen et al. (2021) demonstrated, disfluency features in conversational speech serve as noninvasive biomarkers of moderate-stage Alzheimer's disease, revealing significant differences between AD and age-matched non-AD participants. Modern ASR systems often treat disfluencies as noise and remove them during post-processing,[1] obscuring potentially meaningful details in the transcript (Dinkar, 2022).

During the last couple of years, there have been notable advances in automatic speech recognition for Scandinavian languages, for example Norwegian,[2] and, more importantly for our study, ASR models trained on Swedish data are now readily available (Vesterbacka et al., 2025; Li et al., 2025) – see Section 3.4.

However, such models continue to exhibit substantial performance gaps across a range of neuropsychological assessment settings, largely because they are not trained on acoustically challenging recording conditions, nor on stylistic variability in naturalistic speech production, or speaker characteristics commonly encountered in mental health contexts and early cognitive decline or dialects. Kokkinakis et al. (2025) reported a much higher WER (0.265) on a picture description task ("Cookie Theft") and a substantially lower WER (0.033) on a reading-aloud task, which imposes fewer demands on spontaneous speech production. Such results indicate that ASR systems *can* achieve near-human transcription accuracy on controlled reading tasks, while more spontaneous, cognitively demanding speech—such as narrative or descriptive tasks—often result in higher error rates and reduced reliability for downstream analyses.

[1] Short words may be omitted for a variety of reasons, including insufficient or noisy acoustic evidence, the merging or normalization of tokens during transcription, and the tendency of language models to favor paraphrased outputs that exclude short function words when these result in more probable overall sequences.

[2] Available from: https://huggingface.co/NbAiLab/nb-whisper-small-beta.

3. Participants, Dataset, Format and Technical Details

Originally, the subset of the participants used in the current study were recruited from the *Gothenburg MCI study*, a longitudinal study investigating dementia disorders in patients seeking medical care at a memory clinic (Wallin et al., 2016). The data analyzed here consist of recordings of spontaneous speech collected as part of the separate research project "Linguistic and extra-linguistic parameters for early detection of cognitive impairment" (Riksbankens Jubileumsfond, NHS 14-1761:1 - 2016-2020).

The audio recordings were collected in a relatively controlled environment and feature native Swedish speakers as their first language. The study was approved by ethical approval (reference number: 206–16, 2016; T021-18) issued by the regional ethical review board in Gothenburg, Sweden. Participants were informed that they could withdraw their participation at any time. All data were coded and anonymized. For the audio capture of the task, a Zoom H4n Handy recorder was used, and the resulting audio files were saved and stored as uncompressed audio in .wav-format 44.1 kHz with 16-bit resolution. A speech pathologist and computational linguist were present during the recording sessions, providing all subjects with identical instructions according to a predefined protocol (cf. Section 3.2). The audio recordings were manually transcribed by a professional transcription company using a standardized procedure and clearly specified guidelines to ensure consistency and accuracy. Manual transcribers added full stops at the end of sentences because the same data should be usable for syntactic parsing, which requires explicit sentence boundary marking. All data were stored within a secure university-managed infrastructure, using a information security class 3 platform based on Nextcloud,[3] in compliance with institutional data protection requirements.

3.1. Demographic and Dataset Characteristics

The study sample consisted of speech recordings from 30 participants in the aforementioned study, drawn from *Västra Götaland County* in Sweden. The participants ranged in age from 57 to 78 years (M = 68.9, SD $\approx$ 5.37). The sample included an equal percentage of female/male participants and also equally distributed across the three categories:

[3] https://nextcloud.com.

	$HC_{n=10}$	$MCI_{n=10}$	$SCI_{n=10}$
Females	5	5	5
Males	5	5	5
Mean Age	71.2	68,4	67
Mean Years of Education	13.5	12.8	16.3
Mean Recording Duration	158.3s	130.8s	194s
Mean Token Number	402.5	294,3	426.4
Number of Tokens with Disfluencies	93	86	61
Type-Token Ratio	4.30	3.83	3.82

Table 1: Demographic information for the three groups of participants.

Healthy Controls, HC,[4] *Mild Cognitive Impairment, MCI,*[5] and *Subjective Cognitive Impairment, SCI.*[6] Details of demographic information for the three groups of participants are shown in Table 1.

In the table, Type–Token Ratio (TTR) is a measure of lexical diversity, defined as the number of unique words (types) divided by the total number of words (tokens) in a text or speech sample. In our sample, the TTR is between 4.30 (higher for the healthy group) and 3.82, which is typical for spontaneous informal speech. A higher TTR implies greater lexical diversity, which is often desirable in speech. However, no statistical significance claims can be made due to the small sample size, and the measurements in Table 1 are just given as a reference.

3.2. "Trip to Stockholm": a Swedish Spoken Discourse Task for ASR Evaluation

The spontaneous language material analyzed in the present study was derived from a spoken discourse task modeled on the "Trip to New York" (Harris et al., 2008; Fleming and Harris, 2008). For the purposes of this project, the task was adapted to "Trip to Stockholm" (Antonsson et al., 2021). The use of this complex discourse task has shown the potential to differentiate adults who are normally aging cognitively from those with MCI (Fleming, 2014), which is an important task of our research. Participants were asked to describe orally how they would plan and carry out a trip to Stockholm, following a short series of instructions about the planning of a two week trip:

> *Now you are going to do a task where you are asked to think and plan aloud. Imagine that you are going on a vacation a week from now. You are traveling to Stockholm for a 2-week stay. Think about all you will have to do to get ready to go, such as how you will get there, what you will bring, and what you will do. I want you to tell me all of your plans until I ask you to stop after about 4 to 5 min.*

If participants did not spontaneously include certain information in their narratives, brief follow-up prompts were provided (e.g., "Who will take care of your mail?" or "What will you bring on your trip?"). The task was designed to elicit connected, naturalistic speech requiring conceptual and semantic elaboration related to the cognitive-linguistic schema for travel (Harris et al., 2008). Due to its cognitive and linguistic complexity, the task has been suggested to be sensitive to subtle deficits in individuals with brain injury, engaging executive functions such as initiation, planning, temporal organization, and flexibility, as well as semantic, episodic, and working memory processes.

3.3. Pre-processing

Common disfluency features in Swedish, such as *nonlexical fillers* ”hm”, ”eh” and *vocalizations* "haha", as well as *false starts*, were *not* omitted. Word fragments, e.g., *‘[...] inte betal- beställa något hotell [...]’* (lit. [...] not *pay-* to book a hotel [...]) were transcribed as complete words whenever the intended word could be reliably identified by the manual transcriber; if not, the transcription preserved the original partial or interrupted form; as in *‘[...] när jag har txxx jag tycker [...]’* (lit. [...]’ when I have *txxx*, I think [...]).

Numerical data as well as occurrences of URLs, were rendered in full; for example, "E4"[7] was transcribed as *e-four*; and *bookings.com* as three tokens *bookings punct com*.

In the evaluation all tokens were converted to lowercase and punctuation marks were removed. The Python package `werpy` was used for text normalization.[8] The aim of these normalizations was to

[4] *Healthy Controls* are participants who do not exhibit the condition under investigation and serve as a baseline or comparison group against affected individuals.

[5] *Mild Cognitive Impairment* is defined as a transitional stage of cognitive decline that lies between the alterations associated with normal aging and the deficits that satisfy the diagnostic criteria for clinical dementia (Petersen et al., 2014; Albert et al., 2011).

[6] *Subjective Cognitive Impairment* refers to a self-perceived, persistent decline in one or more cognitive domains over time, occurring in the absence of objectively measurable deficits (Jessen et al., 2007).

[7] "E4" is a major European route (motorway/highway).

[8] https://pypi.org/project/werpy/.

improve the accuracy when matching the transcription with the gold data since difference in phrase segmentation and inconsistencies in using upper-/lowercase letters can have detrimental effects on the evaluation.

3.4. Models and Metrics

Most modern ASR systems are based on OpenAI's Whisper (Radford et al., 2022; OpenAI, 2022), which uses a sequence-to-sequence transformer architecture. Audio is converted to a log-Mel spectrogram, encoded, and then decoded autoregressively into text tokens (words, subwords, or punctuation). Processing occurs in independent segments, which are later combined, allowing efficient transcription but occasionally producing local inconsistencies.

In the present study, ASR transcriptions were generated using locally deployed versions of three publicly available models and variants:

- OpenAI Whisper is an ASR model trained in more than 680,000 hours of multilingual, multitask audio data, designed to support robust transcription and translation across a wide range of languages and recording environments (Radford et al., 2022).[9]
- the Swedish National Library's KB-Whisper, is based on OpenAI's Whisper architecture but trained on over 50,000 hours of Swedish audio (Vesterbacka et al., 2025).[10] The training corpus included TV broadcasts, parliamentary debates, and dialectal recordings, yielding substantially improved accuracy for Swedish speech compared to the original OpenAI model.
- Stable-TS, a variant of OpenAI Whisper, is an open-source timestamp refinement and alignment layer for Whisper-based ASR. It post-processes OpenAI Whisper model output by re-aligning text tokens to audio using forced-alignment and smoothing heuristics.[11]

All three systems were used in their `Faster-Whisper` implementation[12] which utilizes the `CTranslate2` library,[13] a fast inference engine for Transformer models.

We used four size versions of each model: *tiny*, *small*, *medium*, and *large*. Each model configuration, model type, and size, was also assessed using three primary evaluation metrics:

- Word Error Rate (WER): which quantifies the proportion of words incorrectly predicted by a model, accounting for substitutions, deletions, and insertions relative to the reference transcription; that is, the minimum edit distance between a transcript and the reference (ground truth), expressing the proportion of errors relative to the total number of words. The WER metric typically ranges from 0 to 1, where 0 indicates that the compared pieces of text are exactly identical, and 1 (or larger) indicates that they are completely different with no similarity. A WER of 0.8 means that there is an error rate of 80% for the compared sentences.
- Bilingual Evaluation Understudy (BLEU): which measures the n-gram overlap between the predicted and reference transcriptions; (Papineni et al., 2002). BLEU computes a value between 0 and 1, where 1 corresponds to perfect agreement between the prediction and the gold standard. Although BLEU was originally developed for machine translation evaluation, it has also been applied to speech-to-text output by comparing the generated transcript with a reference text. However, this metric does not fully capture recognition errors and should therefore be interpreted with caution.
- Google-BLEU (GLEU): a measure intended to overcome limitations in BLEU score calculations and are better suited for sentence level comparisons GLEU balances precision and recall over 1-4 n-grams between predicted and reference transcriptions. (Mutton et al., 2007).

We used the BLEU and GLEU implementations from NLTK[14] as well as the WER implementation from the `werpy` package.

The study also considers a range of the hyperparameter *temperature* settings (0, 0.25, 0.5, 0.75, and 1), which control the degree of randomness during decoding. Lower temperature values lead to more deterministic and stable transcriptions, whereas higher values introduce increased variability in the generated output, potentially capturing alternative word choices at the cost of consistency, i.e. increased error rates (cf. Table 2).

[9] https://huggingface.co/openai/whisper-large-v3.
[10] https://huggingface.co/KBLab/kb-whisper-large.
[11] https://github.com/jianfch/stable-ts.
[12] https://github.com/SYSTRAN/faster-whisper.
[13] https://github.com/OpenNMT/CTranslate2/.
[14] https://www.nltk.org/.

4. Evaluation and Analysis

The three ASR models are assessed against the reference transcripts, using a version without any punctuation markings. The transcripts were pre-processed according to the previous description (Section 3.3). Numerical tokens were converted to text (e.g. "4" to "four") and all transcriptions were transformed to lower-case. We evaluated the performance of each ASR model using the previously described metrics, WER, BLEU, and GLEU, and the five temperature hyperparameters.

Swedish short discourse adverbs and function words (2-3 characters long), such as *väl* (lit. “well”) and *ju* (a discourse particle roughly meaning “as you know” or “after all”) are often dropped in transcriptions, probably because of acoustic ambiguity and language model bias. For example, *ska väl gå* becomes *ska gå* (lit. “it should work”) or *’a så har de ju vasamuseet’* becomes *’och så har de vasamuseet’* (lit. “and then they have the Vasa Museum, of course”); see also footnote 1. Similar behavior is observed in multiword function words such as *i och med* (“given that” or “due to”); *nu är jag ju bortskämd i och med att jag har en hustru* (lit. “I am, of course, spoiled, given that I have a wife”) to *nu är jag bortskön att jag har en hustru* (lit. “Now I am ‘bortskön’ that I have a wife”); note also the wrong annotation of *bortskämd* to *bortskön*.

Some other discrepancies between near-verbatim manual transcription and the models’ output, as well as between orthographic and phonetic-near transcriptions can be explained by:

(i) homonym-phonetic confusion (’å’ sometimes ’och’ [and]);

(ii) occasional cases in which the orthographic transcription incorporated manually asserted phonetic symbols (e.g., ’n:u’ [now] and ’- -’ longer pauses), reflecting a partial convergence with phonetic-level representation;

(iii) phonological deviation ’å slå sej’ (lit. ’och slå sig’) – [and beat themselves].

More importantly, *ASR artefacts*, such as the outputted word “lagoda”, may result from word concatenation, erroneous normalization, or phonetic approximation under conditions of rapid or indistinct Swedish speech; in this instance, the form most likely corresponds to a misrecognition of the proper name “Agoda”, a travel agency.

Finally, *overregularization*, when a grammatically valid rule is applied in an inappropriate linguistic context, is also observed in such cases models apply the regular Swedish plural suffix (here ’-ar’) to nouns that exhibit zero plural marking in standard Swedish; e.g. ’fyra lamm’ (lit. four lambs) to ’fyra lammar’; or ’skjortor’ (lit. shirts) as ’skjortar’.

4.1. Performance Results

The evaluation results are shown in Table 2. As can be seen in this table there are systematic differences in performance depending both on the type of model and the size configuration. Larger configurations generally yield higher accuracy, though gains are model-dependent, and some smaller configurations achieve competitive results, suggesting potential trade-offs between computational cost and performance.

In all tasks, as expected, the large Swedish *KB-Whisper* model performed overall best for all three metrics, with a temperature set to 0.5. In fact, 0.147 was the best overall WER value, 0.927 the best overall value for BLEU, and 0.926 the best overall GLEU value, regardless of the model. Lower WER values generally indicate transcripts that more closely approximate the reference text and are therefore more likely to be understandable. While high BLEU and GLEU values indicate that the generated transcripts closely match the reference text in terms of word choice and local phrase structure. Notably, the lowest WER values for *OpenAI* and *Stable-TS* were the medium models with temperature=0, 0.202 and 0.199, respectively. As an outlier the *OpenAI-large* model started repeating a part of the output for one of the example over and over again which resulted in very bad scores. We are not sure about the source of this problem but were able to reproduce with the same input and parameters.

The results show consistent differences in the processing/transcription time between model sizes and resource settings.[15] For both KB-Whisper and OpenAI Whisper, the *large* models require substantially longer total runtime than the *tiny* variants (approximately 13–14 minutes vs. just over 8 minutes), with average per-instance processing times of around 5.5 seconds for *large* and 3.3 seconds for *tiny*. In contrast, the Stable-TS setting is markedly more computationally demanding, with runtimes increasing by a factor of approximately four for all models, particularly for the *large* configuration (56 minutes total, 22.5 seconds on average).[16] Despite these differences, the relative efficiency gap between *large* and *tiny* remains stable across settings, indicating that model size consistently impacts computational cost, while the choice of resource configuration has a much stronger effect on overall runtime.

[15] All data has been transcribed using the CUDA implementation of Faster Whisper on a server equipped with a NVIDIA GeForce RTX 3060 (12 GB RAM).

[16] The length of the audio recordings is ca 80 min, all transcriptions were substantially faster than real-time.

			temp=0	temp=0.25	temp=0.5	temp=0.75	temp=1
KB-Whisper	large	WER	0.149	0.153	**0.147***	0.163	0.189
		BLEU	0.926	0.922	**0.927***	0.916	0.903
		GLEU	0.924	0.920	**0.926***	0.914	0.900
	medium	WER	**0.206**	**0.206**	0.207	0.216	0.219
		BLEU	**0.879**	0.881	0.877	0.869	0.870
		GLEU	0.874	**0.876**	0.872	0.863	0.864
	small	WER	**0.171**	0.173	0.179	0.182	0.196
		BLEU	**0.913**	0.909	0.910	0.903	0.895
		GLEU	**0.912**	0.907	0.908	0.901	0.891
	tiny	WER	0.565	0.683	0.583	**0.244**	0.279
		BLEU	0.571	0.499	0.534	**0.885**	0.867
		GLEU	0.505	0.415	0.447	**0.882**	0.863
OpenAI	large	WER	0.674	0.273	0.658	**0.224**	0.273
		BLEU	0.837	0.862	0.799	**0.875**	0.847
		GLEU	0.835	0.859	0.796	**0.869**	0.840
	medium	WER	**0.202***	0.218	0.227	0.245	0.364
		BLEU	**0.904***	0.893	0.883	0.868	0.790
		GLEU	**0.902***	0.890	0.878	0.863	0.776
	small	WER	**0.253**	0.290	0.298	0.317	0.476
		BLEU	**0.880**	0.849	0.851	0.841	0.729
		GLEU	**0.876**	0.842	0.845	0.834	0.708
	tiny	WER	**0.476**	0.646	0.584	0.605	0.774
		BLEU	**0.795**	0.727	0.759	0.726	0.544
		GLEU	**0.785**	0.716	0.748	0.706	0.491
Stable-TS	large	WER	0.220	0.274	0.310	**0.210**	0.268
		BLEU	**0.893**	0.867	0.855	0.886	0.855
		GLEU	**0.890**	0.865	0.852	0.882	0.849
	medium	WER	**0.199***	0.234	0.224	0.237	0.365
		BLEU	**0.904***	0.874	0.886	0.879	0.788
		GLEU	**0.902***	0.869	0.883	0.875	0.770
	small	WER	**0.259**	0.283	0.292	0.326	0.490
		BLEU	**0.874**	0.855	0.852	0.834	0.716
		GLEU	**0.870**	0.849	0.845	0.826	0.692
	tiny	WER	**0.510**	0.537	0.572	0.603	0.772
		BLEU	**0.770**	0.760	0.739	0.728	0.564
		GLEU	**0.759**	0.746	0.723	0.708	0.521

Table 2: Evaluation results for the three main models, *KB-Whisper*; *OpenAI*; *Stable-TS*, comparing performance across the four size configurations (*tiny*, *small*, *medium* and *large*) and five temperature values for each model architecture (*0*, *0.25*, *0.5*, *0.75* and *1*). The best results for each temperature value are marked in bold and the overall best result for each model is marked with an asterisk (*).

5. Discussion and Future Work

Despite major advances in ASR and claims of near-human precision, evaluations in domains such as Higher Education lectures reveal substantial variability and reduced reliability, particularly for streaming applications (Kuhn et al., 2024). Although our study is subject to limitations, such as the inclusion of only 30 participants, the findings may nonetheless provide valuable insights for downstream applications. In particular, when combined with careful preprocessing and quality control, these approaches can support automated cognitive evaluation and the monitoring of language-related decline, especially in large-scale evaluations involving population-level cohorts.

We strongly believe that the WER (and other metrics) on this kind of data can be improved through several practical strategies, including enhancing audio quality, employing domain-specific language models, applying post-processing corrections, or retraining the system with additional in-domain speech data. These approaches are primarily applicable in future data collection scenarios and can involve recordings of individuals with mild or severe cognitive impairments, although the collection of such data necessarily requires careful ethical consideration.

By contrast, we want to explore *prompting*. We can apply this method directly to existing recordings without the need of additional data acquisition or training. Further assessment of the robustness of this framework, together with evaluation of its performance, constitutes a primary focus of future work. Prompts for Whisper models are used to stitch together multiple audio segments, Whisper is using a sliding audio context window of 30 seconds.[17] However, giving an initial prompt can even steer the model output, providing spelling and output formatting hints.[18]

In addition, future investigations will extend the evaluation to newly released and updated Whisper-inspired model versions,[19] enabling a more comprehensive comparison.

On the research infrastructure side we work on extending the tooling provided by Språkbanken Text. Whisper-based transcriptions (Språkbanken Text, 2025c) are already available in the Sparv pipeline (Språkbanken Text, 2025d) and consequently also in the Mink platform (Språkbanken Text, 2025a,b). However, the feature set is currently too limited to conduct our experiment within this framework. We plan to add the missing features required by our experiments, allowing us as well as other researchers to reproducibly repeat this and similar experiments using the Sparv (Språkbanken Text, 2025d) pipeline. The full Whisper toolbox will be made available within Mink. More details about Mink and Sparv can be found in Forsberg et al. (2025). The current code for the experiment and evaluation is also available on Github under a free and open license.[20]

[17] https://github.com/openai/whisper.

[18] https://developers.openai.com/cookbook/examples/whisper_prompting_guide.

[19] Other models *not* considered in the study include https://huggingface.co/birgermoell/whisper-small-sv-bm and the https://github.com/m-bain/whisperX.

[20] https://github.com/spraakbanken/Whisper-experiment/.

6. Conclusion

This study evaluates state-of-the-art automatic speech recognition (ASR) models for the full-scale automatic transcription of a Swedish discourse narrative task. The population in focus includes individuals with early signs of cognitive impairment. ASR provides a scalable and automated method for analyzing spoken responses in cognitive assessments. However, spontaneous speech—often characterized by disfluencies, hesitations, and complex syntactic structures—remains more challenging than controlled reading tasks, influencing recognition accuracy and downstream analysis such as automatic scoring or classification tasks.

We focus on Word Error Rate (WER) because it directly measures word-level transcription accuracy, which is the primary objective of this work. WER reflects human correction effort, penalizes substitutions, deletions, and insertions symmetrically, and discourages over-generation and hallucination. Although WER does not capture semantic equivalence, it is a deliberate and widely accepted choice to evaluate transcription fidelity and ensures comparability with previous ASR literature. In addition to WER, we report metrics in BLEU and GLUE-style to capture complementary aspects of ASR output quality. WER measures exact word-level transcription fidelity, whereas BLEU reflects local phrase consistency and fluency by providing partial credit for benign lexical variations. GLUE-style metrics assess semantic equivalence, enabling us to distinguish meaning-preserving deviations from semantically harmful errors. Together, these metrics offer a more complete evaluation while retaining WER as the primary measure of transcription accuracy.

This evaluation therefore offers practical guidance for selecting ASR models suited to Swedish-language clinical and research applications, balancing transcription quality with robustness to natural speech variations. By mapping the performance of the model to the demands of specific tasks, we outline a framework for integrating AI transcription into screening workflows. The findings underscore the importance of task-sensitive model evaluation and support the development of automated tools and platforms for cognitive evaluation.

7. Limitations

The dataset is limited by the small sample size, comprising only 30 participants from the same geographical area and with a comparable age and level of education. This restricts the generalizability of the findings and calls for caution when interpreting the results.

AI models can produce confusing words (or sentences) by mistaking homonyms or hallucinating text, particularly in cases where contextual cues are weak and the model must rely on uncertain predictions (Koenecke et al., 2024). Moreover, evaluation metrics such as WER — though simple and easy to compute — have been criticised for not capturing text understanding and for correlating only weakly with human judgments of transcript quality (Just et al., 2025; Phukon et al., 2025). Still, WER remains one of the most widely used and practical metrics for evaluating ASR systems, as their verbatim outputs lend themselves to word-by-word comparison.

8. Ethical Considerations

During the experiments, we ensured that no private or personally identifiable information—such as participants' names and health data — was disclosed or processed outside the local environment. To minimize privacy risks and maintain full control over the data, all experiments were conducted exclusively using locally installed open-source models. This approach ensured that no data were transmitted to external servers or third-party services, thereby complying with data protection and ethical research standards.

9. Disclosure

The authors used the digital assistant platform *ChatGPT Edu version 5.2*, to support limited aspects of the writing process, specifically grammar, morphological refinement, and related checks (e.g., spelling verification and typographical correction). All conceptual contributions, analyses, interpretations, and conclusions are solely the authors' own, and no generative tool was used to produce or modify empirical data, figures, or results.

10. Acknowledgments

The research presented here was supported by the Swedish Research Council (grant number 2025-00765), the Swedish national research infrastructure Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (grant numbers 2017-00626 and 2023-00161), the Huminfra, the Swedish national infrastructure for the Humanities, funded by the Swedish Research Council and the consortium nodes (grant numbers 2021-00176 and 2023-00171), as well as by the The Swedish Parkinson Foundation (Parkinsonfonden).

11. Bibliographical References

Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J. Snyder, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. 2011. The diagnosis of mild cognitive impairment due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):270–279.

Malin Antonsson, Kristina Lundholm Fors, Marie Eckerström, and Dimitrios Kokkinakis. 2021. Using a discourse task to explore semantic ability in persons with cognitive impairment. *Frontiers Aging Neuroscience*, 12.

Eleonora J Beier, Suphasiree Chantavarin, and Fernanda Ferreira. 2023. Do disfluencies increase with age? evidence from a sequential corpus study of disfluencies. *Psychology and Aging*, 38(3):203–218.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Ronald Cumbal, Birger Moëll, Jose Lopes, and Engwall Olof. 2024. "you don't understand me!": Comparing asr results for l1 and l2 speakers of swedish. *Computing Research Repository*, arXiv:2405.13379v1. Version 1.

Tanvi Dinkar. 2022. Computational models of disfluencies : fillers and discourse markers in spoken language understanding. *Computer science. Institut Polytechnique de Paris*, NNT: 2022IPPAT001.

Valarie Fleming and Joyce L. Harris. 2008. Complex discourse production in mild cognitive impairment: Detecting subtle changes. *Aphasiology*, 22:792–740.

Valarie B. Fleming. 2014. Early detection of cognitive-linguistic change associated with mild cognitive impairment. *Communication Disorders Quarterly*, 35:146–157.

Markus Forsberg, Dana Dannélls, Lars Borin, and Aleksandrs Berdicevskis. 2025. *Background: Språkbanken Text*, chapter 9. De Gruyter.

Joyce L. Harris, Swathi Kiran, Thomas P. Marquardt, and Valarie B. Fleming. 2008. Communication wellness check-up©: Age-related changes in communicative abilities. *Aphasiology*, 22:813–825.

Frank Jessen, Birgitt Wiese, Gabriela Cvetanovska, Angela Fuchs, Hanna Kaduszkiewicz, Heike Kölsch, Tobias Luck, Edelgard Mösch, Michael Pentzek, Steffi G Riedel-Heller, Jochen Werle, Siegfried Weyerer, Thomas Zimmermann, Wolfgang Maier, and Horst Bickel. 2007. Patterns of subjective memory impairment in the elderly: association with memory performance. *Psychol Med.*, 37(12):1753–62.

Yue Jiang and Xufei An. 2025. Speech differences between aged women with and without early alzheimer's disease: linguistic indicators of cognitive decline. *Acta Psychologica*, 256.

Sandra Anna Just, Brita Elvevåg, Shrankhla Pandey, Ivan Nenchev, Anna-Lena Bröcker, Christiane Montag, and Sarah E Morgan. 2025. Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders. *Psychiatry Research*, 352.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1672–1681, New York, NY, USA. Association for Computing Machinery.

Dimitrios Kokkinakis, Herbert Lange, and Ricardo Muñoz Sánchez. 2025. Evaluating speech-to-text models for swedish neuropsychological assessments: a comparative study across task types and models. In *Proceedings of the 35th Alzheimer Europe conference "Connecting science and communities: The future of dementia care*, Bologna, Italy.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2024. Measuring the accuracy of automatic speech recognition solutions. *ACM Trans. Access. Comput.*, 16(4).

Changye Li, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2024. Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of Biomedical Informatics*, 150.

Zirui Li, Jens Edlund, Yicheng Gu, Nhan Phan, Lauri Juvela, and Mikko Kurimo. 2025. Nordparl-tts: Finnish and swedish tts dataset from parliament speech.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, 3.

OpenAI. 2022. Whisper: Robust speech recognition model. `https://github.com/openai/whisper/blob/main/README.md`. Accessed: 2026-01-16.

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 on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Ronald C Petersen, Barbara Caracciolo, Carol Brayne, Serge Gauthier, Vesna Jelic, and Laura Fratiglioni. 2014. Mild cognitive impairment: a concept in evolution. *J Intern Med.*, 275:214–228.

Bornali Phukon, Xiuwen Zheng, and Mark Hasegawa-Johnson. 2025. Aligning asr evaluation with human and llm judgments: Intelligibility metrics using phonetic, semantic, and nli approaches. In *Proceedings of the 26th Interspeech*, pages 5708–5712, Rotterdam, The Netherlands. Association for Computational Linguistics.

Aurélie Pistono, Jérémie Pariente, and Mélanie Jucla. 2024. Disfluency patterns in alzheimer's disease and frontotemporal lobar degeneration. *Clinical Linguistics & Phonetics*, 38(4):345–358.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Farzana Shahla, Deshpande Ashwin, and Natalie Parde. 2022. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 37–48, Dublin, Ireland. Association for Computational Linguistics.

Leonora Vesterbacka, Faton Rekathati, Robin Kurtz, Justyna Sikora, and Agnes Toftgård. 2025. Swedish whispers; leveraging a massive speech corpus for swedish speech recognition. In *Proceedings of the Interspeech*, pages 758–762, Rotterdam, The Netherlands.

Anders Wallin, Arto Nordlund, Michael Jonsson, Karin Lind, Åke Edman, Mattias Göthlin, Jacob Stålhammar, Marie Eckerström, Silke Kern, Anne Börjesson-Hanson, Mårten Carlsson, Erik Olsson, Henrik Zetterberg, Kaj Blennow, Johan Svensson, Annika Öhrfelt, Maria Bjerke, Sindre Rolstad, and Carl Eckerström. 2016. The gothenburg mci study: Design and distribution of alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *J Cereb Blood Flow Metab.*, 36(1):114–131.

12. Language Resource References

Språkbanken Text. 2025a. *Mink*. Språkbanken Text. https://spraakbanken.gu.se/mink/.

Språkbanken Text. 2025b. *sbx-swe-mink_analyses*. Språkbanken Text. https://doi.org/10.23695/r5q1-xa67.

Språkbanken Text. 2025c. *sbx-swe-speech2text-transformers-kb_whisper_mp3*. Språkbanken Text. https://doi.org/10.23695/6nry-qr23.

Språkbanken Text. 2025d. *Sparv*. Språkbanken Text. https://spraakbanken.gu.se/sparv/.

ALBA: An Automated Framework for Benchmarking Clinical Language Biomarkers against Standardized Corpora

Charalambos Themistocleous, Brielle C. Stark
[1]University of Oslo, 0371 Oslo, Norway
[2]Indiana University Bloomington, 2631 E Discovery Parkway, Bloomington, IN 47408, USA
[1]charalth@uio.no, [2]bcstark@iu.edu

Abstract

Patients with diverse neurocognitive conditions frequently exhibit measurable language deficits that serve as biomarkers for differential diagnosis and therapy decision making. Discourse analysis can offer reliable ecological measures of human communication, yet manual discourse analysis is cumbersome. Recent advances in automated analysis software provide quick and easy extraction of raw language metrics in the clinic. Nevertheless, transforming these measures into actionable clinical insights remains a significant challenge. The aim of this paper is to present the Automated Language Biomarker Application (ALBA), an integrated framework developed within the Open Brain AI ecosystem to bridge the gap between feature extraction and clinical interpretation. ALBA provides clinicians with a robust statistical infrastructure to benchmark individual patient measures against standardized, large-scale clinical corpora. By utilizing a shared elicitation and processing pipeline, the application ensures that user-provided data are directly comparable to population norms for conditions including Aphasia, Mild Cognitive Impairment (MCI), Dementia, and other neurological conditions. The system implements adaptive statistical logic, employing one-sample t-tests and robust non-parametric alternatives to provide real-time significance testing and dynamic visualizations (box, bar, and violin plots). By automating the comparison of "Language Signatures" against healthy controls and specific clinical phenotypes, ALBA facilitates rapid, evidence-based decision-making in both research and rehabilitation contexts.

Keywords: Language Biomarkers, Open Brain AI, TalkBank, Clinical Discourse Analysis

1. Introduction

Clinical discourse analysis has a long history in assessment of neurogenic language, especially aphasia, because language-in-use, such as during picture sequence description or narrative retells, often provides a more nuanced and complex picture of a person's linguistic ability than isolated tests of language, such as confrontation picture naming. While clinical discourse analysis evaluates linguistic, propositional, macrostructural, and pragmatic aspects of discourse, many neurogenic populations struggle with foundational linguistic components. As a result, language biomarkers have tended to be derived from quantitative language measures linked to specific clinical or developmental conditions. Further, in practice, clinicians often extract such measures, including part-of-speech counts, lexical ratios, and word-frequency distributions, to systematically assess and objectively score an individual's communicative efficacy (Bryant et al., 2016).

However, interpreting these findings remains challenging in the absence of reference corpora or standardized norms to facilitate comparative analysis. To bridge this gap, three primary objectives must be met: the establishment of gold-standard metrics, the adoption of shared elicitation methodologies, and the provision of accessible tools for the rapid comparison of results to support informed clinical decision-making.

Although numerous research studies have reported various linguistic measures (Themistocleous and Stark, 2026; Varkanitsa et al., 2023; Kiran et al., 2019), the field currently lacks a unified framework that enables clinicians to generate patient-specific data and compare it with standardized measures derived through identical methodologies. As the need for clinical discourse analysis and its importance in eliciting rich and ecological measures of human communication compared to other clinical tasks becomes more pressing, providing such a unified framework is an exceedingly important need.

This paper presents Automated Language Biomarker Application (ALBA), a web-based tool for producing and comparing clinical research outcomes with standardized measures. ALBA has been developed to serve as a data resource by facilitating the presentation and comparison of automatic measures generated by Open Brain AI, a clinical research platform and computational tool designed for language assessment and analysis (Themistocleous, 2024). These measures originate from both clinical tasks (e.g., cookie theft, Cinderella story, story-telling, and story-retelling) as well as from large-scale text corpora. A key advantage of ALBA is its flexibility: new language measures can be easily added or updated to accommodate emerging clinical conditions and evolving research needs.

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 34–40
12 May 2026.

2. Previous Research

Developmental or acquired neurocognitive conditions can disrupt language and communication in complex and heterogeneous ways, with impairments often spanning expressive domains such as phonology, morphology, syntax, semantics, and lexical access (Themistocleous and Stark, 2026; Varkanitsa et al., 2023; Kiran et al., 2019). While some aspects of linguistic breakdown can be partially predicted by lesion location or severity, the relationship is far from deterministic due to the distributed and dynamic nature of language networks in the brain. Discourse-level language analysis offers a unique window into these impairments, capturing subtle disruptions in coherence, cohesion, informativeness, and pragmatic appropriateness that are often missed by more constrained or modular assessments.

Through systematic quantification of expressive language—particularly at the discourse level—automated language measures can function as robust biomarkers, helping to characterize, differentiate, and subtype various neurological conditions, including left and right hemisphere stroke, traumatic brain injury, mild cognitive impairment, and dementia. These measures not only distinguish clinical populations from healthy controls but also provide insight into the underlying cognitive and communicative mechanisms affected in each condition (Themistocleous and Stark, 2026).

To enable differential diagnosis and inform treatment planning, language measures must be interpreted against well-characterized reference populations or normative data, which provide essential context for distinguishing clinical profiles. Normative reference data are especially valuable in identifying subtle but clinically meaningful language changes in individuals with the mildest forms of aphasia (e.g., latent aphasia), traumatic brain injury, right hemisphere disorder, mild cognitive impairment, or the earliest stages of dementia—conditions where language impairments may not be immediately obvious but are crucial to differentiate from typical aging. While normative comparisons may be less critical for diagnosing more overt or "frank" aphasias, they still provide meaningful context for characterizing the specific pattern and severity of impairment, guiding tailored treatment approaches, and establishing a baseline for tracking individual progress.

For clinicians, access to normative data can enhance decision-making in several key ways. (1) To characterize the specific pattern and severity of impairment, detailed language measures—such as lexical diversity, syntactic complexity, or informativeness—can help phenotype aphasia presentations beyond broad classifications like Broca's or Wernicke's aphasia. For instance, two individuals with anomic aphasia may show similar naming deficits, but one may produce overly vague narratives with limited cohesion, while another struggles with syntactic formulation—distinctions that are only visible through discourse-level profiling against normative benchmarks. (2) In guiding tailored treatment approaches, knowing which language domains are disproportionately affected relative to healthy controls can help clinicians prioritize intervention targets. For example, if a person with right hemisphere damage demonstrates relatively preserved syntax but poor global coherence, therapy can focus more on narrative structuring and pragmatic use. (3) To establish a baseline, quantified language data at intake provide a reference point for monitoring individual progress over time, allowing clinicians to track meaningful change in discourse abilities, even if those changes do not shift the person's broad diagnostic category.

2.1. Traditional Discourse Analysis: Time Consuming, Resource Intensive, Lacking in Tools

Historically, discourse analysis in clinical and educational settings has relied on manual annotation, a process requiring granular characterization of *micro-structural* features (linguistic and propositional levels, e.g., lexical and sentence level features) and *macro-structural* properties (planning and pragmatic levels, e.g., global coherence/cohesion, thematic evolution, and topic maintenance) (Paltridge, 2006). Although providing deep qualitative insights, manual analysis is resource-intensive, requiring specialized linguistic expertise and extensive labor for scoring (Hansen et al., 2022; Cruice et al., 2020; Bryant et al., 2018), and often clinicians and researchers cite these as significant barriers preventing them from engaging in discourse analysis. Consequently, traditional diagnostic batteries have often been restricted to narrow elicitations of measures from connected speech productions and a substantial lack of standardization outputs (Stark et al., 2023).

2.2. Computational Tools

To address the limitations of manual scoring, several software frameworks have emerged to automate linguistic feature extraction. The *Computerized Language Analysis (CLAN)* system, part of the TalkBank project, established the industry standard by using the CHAT transcription format to calculate morphosyntactic and lexical diversity measures (MacWhinney). Similarly, the *Systematic Analysis of Language Transcripts (SALT)* has become a clinical staple for speech-language pathologists (SLPs), focusing on standardized metrics like Mean Length

of Utterance (MLU) and error coding (Cunningham and Haley, 2020; Fergadiotis, 2011).

The emergence of neural Natural Language Processing along with end-to-end automated pipelines for text processing are gradually finding their way into the clinic (Tippett et al., 2025). Platforms such as *Open Brain AI (OBAI)* (Themistocleous, 2024) and automated tools like *Batchalign* (Liu et al., 2023) have moved beyond keyword counting to leverage Automatic Speech Recognition (ASR), Neural Morphosyntactic Tagging, and Transdiagnostic Biomarkers. For example, Open Brain AI can now extract hundreds of linguistic biomarkers—ranging from acoustics to semantic coherence–allowing for a "Language Biomarker" that characterizes specific neurodegenerative or developmental conditions. Such automated computational language measures are being to describe the various language domains including and contribute to the automatic patient identificatio, subtyping, and prognosis (Fraser et al., 2014; König et al., 2018; Themistocleous et al., 2021).

Modern automated AI tools typically generate two distinct classes of linguistic measures, each serving a unique function in clinical research. The first class consists of interpretable biomarkers with direct physical or clinical correlates. These measures are associated with specific pathologies; for instance, deficits in function word ratios can indicate agrammatism, while a reduced noun-to-verb ratio may signal anomia (Themistocleous et al., 2020). Because of their overt clinical interpretation, these metrics are easily integrated into diagnostic assessments and used to define specific therapeutic targets.

The second class comprises latent representations, such as high-dimensional word and sentence embeddings (Bengio and Heigold; Mikolov et al., 2013). While these measures—often derived from large-scale transformer models—lack immediate transparency for clinicians, they serve as highly robust predictors in supervised and unsupervised classification tasks. Both types of measures are essential: while interpretable features provide the "why" for clinical intervention, latent embeddings often provide superior accuracy for automated screening and condition subtyping.

2.3. From Extraction to Interpretation

Despite the proliferation of tools designed to *extract* linguistic data from discourse, a critical gap remains in the *interpretation* of the interpretable language biomarkers within a clinical time-frame. Although existing software provide raw counts or ratios, the burden of statistical comparison often falls on the clinician, who must manually reference published norms or reference datasets. Our application, ALBA (`https://openbrainai.com/measures`), is integrated directly within the *Open Brain AI* ecosystem and addresses this limitation by providing a robust statistical infrastructure that bridges the gap between computational measures and clinical decision-making.

First, it utilizes large-scale normative measures for mitigating the impact of individual linguistic idiosyncrasies (Stark and Fukuyama, 2021). Historically, the use of standardized corpora has been foundational to clinical and psychological research. Early efforts relied on general-purpose datasets, such as the *Brown Corpus* in English and the Språkbanken Text to establish baseline word frequencies and lexical expectations (Francis and Kucera, 1982; Forsberg et al., 2025). Large-scale collections are critical because they define the boundaries of "normal" linguistic variation, accounting for the vast diversity in human speech influenced by age, education, and cognitive health. Without large, validated datasets, researchers and clinicians are unable to determine if a patient's performance represents a pathological deficit or simply a point within the tail of typical variation. Language measures from large-scale normative datasets are thus essential for mitigating the impact of individual linguistic idiosyncrasies serving as benchmarking data for the scores elicited in the clinic.

ALBA aggregates measures derived from standardized corpora, offering an interactive environment where practitioners can empirically determine the degree of divergence between their specific patient data and established population norms. Specifically, by benchmarking against HCs clinicians can quantitatively assess whether a patient's linguistic profile falls within a normative range. Furthermore, it allows for differential comparisons across a spectrum of neurological conditions, including LHD, RHD, TBI, MCI, and dementia.

Secondly, ALBA's integration with Open Brain AI ensures methodological consistency by enabling a direct comparison between clinician-elicited measures and standardized benchmarks, both of which are processed through the same computational pipeline. This homogeneity eliminates cross-platform variance and enhances interpretative accuracy substantially.

3. The ALBA User Interface

The interface of the *(ALBA)* application shown in Figure 1 is designed for clinical intuition and research rigor. The architecture is divided into two primary functional zones:

The *Configuration Sidebar (Left Panel)* facilitates the parameterization of the analysis. Users can select the specific elicitation *task* (e.g., the Cinderella story-retell), the *category* of linguistic measures (lexical, phonological, morphological, syntactic, or

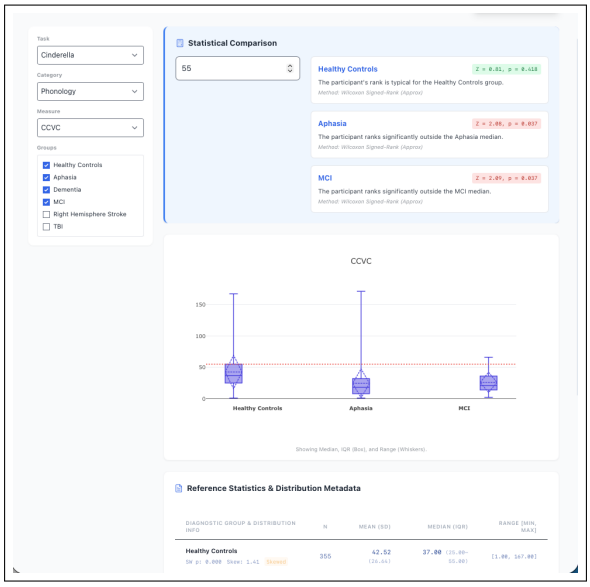


Figure 1: Interface of the Linguistic Biomarkers Application (ALBA). The left hand area allows the selection of the corpus, the category of linguistic measures (e.g., lexical, phonological, morphological, syntactic, semantic) and the language group (healthy controls, patients with different conditions). The main area includes the statistical interface that provides a score with a p-value on whether the measure provided by clinicians is the same or different from the one in the visible corpora based on statistical analysis (see Statistical Methodology). Users can selected different visualization options and view descriptive statistics as a table.

semantic), and the target *clinical cohorts* for comparison (e.g., Healthy Controls vs. various Aphasia subtypes).

The *Analytical Dashboard (Main Panel)* constitutes the central workspace features, an interactive statistical comparison tool. Upon entering a patient's score, the system dynamically calculates and reports a p-value and a formatted APA-style narrative indicating whether the individual's performance significantly diverges from the selected reference corpora (see *Statistical Methodology*).

To support diverse interpretative needs, the main panel offers multiple visualization modalities, including box plots, bar charts with standard deviation error bars, and density-style violin plots. Below the visual representations, a comprehensive Reference Statistics Table provides metadata—including sample sizes, medians, Interquartile Ranges (IQR), and pre-computed normality indicators—ensuring full transparency of the underlying normative data.

4. Methodology

4.1. Data

A flattened relational schema designed for rapid client-side parsing generated from an earlier study (Themistocleous and Stark, 2026) using Open Brain AI from TalkBank data (MacWhinney) is utilized as the primary dataset.

4.2. Statistical Methodology

To maintain high system performance and low latency, ALBA utilizes a hybrid decision engine for statistical test selection. Rather than performing resource-intensive computations on raw transcripts ($N > 600$) at the client level, the system utilizes precomputed distributional metadata. These data correspond to a unique combination of *Task*, *Diagnosis*, *Linguistic Variable* with eleven foundational metrics. *Central Tendency and Dispersion* (i.e., Mean, Standard Deviation (SD), and Median), *Distribution Geometry* (i.e., Min, Max, and the Interquartile Range (Q_1, Q_3)), *Inferential Metadata* (Precomputed Shapiro-Wilk p-values ($P\_Shapiro$) and Fisher-Pearson skewness coefficients ($Skewness$), which inform the automated `IsNormal` flag).

4.3. Test Selection Criteria

The application evaluates each linguistic variable against two primary assumptions:

1. *Parametric Path:* If the reference distribution is pre-validated as normal via the Shapiro-Wilk test ($p > .05$) or meets the Central Limit Theorem criteria ($n \geq 30$ with low skewness), a *one-sample t-test* is performed.

2. *Non-Parametric Path:* If the distribution exhibits significant skewness ($|\mu - M| > 0.1\mu$) or fails normality testing, the system automatically employs a *one-sample Wilcoxon signed-rank* approximation based on the median and Interquartile Range (IQR).

4.4. Parametric Comparison

When the normality criteria are met, the system calculates a one-sample t-test to determine if the user value significantly differs from the reference mean (μ). The t-statistic is calculated as:

$$t = \frac{\mu - x_{observed\ individual\ score}}{SE} \quad (1)$$

where the Standard Error (SE) is defined as:

$$SE = \frac{\sigma}{\sqrt{n}} \quad (2)$$

The p-value is then derived from the t-distribution with $df = n - 1$ degrees of freedom.

4.5. Non-Parametric Comparison

For the non-normally distributed pathological speech data in the standardized datasets, the system employs a non-parametric approach utilizing the Wilcoxon Signed-Rank test. For a single user-provided observation ($x_{observed\ individual\ score}$), we evaluate the probability of observing a value at least as extreme as $x_{observed\ individual\ score}$ given the reference median (M) and the distribution of ranks. The test statistic W is calculated as:

$$W = \sum_{i=1}^{n} \text{sgn}(x_i - x_{observed\ individual\ score}) \quad (3)$$

In our implementation, the system approximates the p-value by determining the percentile rank of $x_{observed\ individual\ score}$ within the reconstructed distribution of the reference group.

The application automatically generates a clinical explanation of the output following the *Publication Manual of the American Psychological Association* (7th ed.) standards.

4.6. Implementation and Visualization

The statistical engine is implemented using the `jStat` library, allowing for real-time, client-side computation. This ensures low latency and enhances user privacy, as the comparative value x_{user} remains local to the user's browser.

To provide immediate clinical intuition, the system overlays the user's value onto the population distribution using a dynamic reference line across three visualization modes Box Plots, Bar Charts with *Significance Feedback*. The user interface provides a "Significant" (red) or "Similar" (green) status based on an alpha level of $\alpha = 0.05$.

5. Conclusion

In this paper, we presented the Automated Language Biomarker Application (ALBA), an integrated statistical interface designed to bridge the gap between automated linguistic feature extraction and clinical interpretation. By situating ALBA within the Open Brain AI ecosystem, we have created an accessible workflow that converts isolated language discourse measures into actionable "Language Biomarkers." The contribution of this work is three-fold. First, we demonstrate how modern AI pipelines can be coupled with clinical resources like *TalkBank* to automate discourse analysis. Second, ALBA provides quick and easy statistical comparisons of clinical measures, distinguishing between typical linguistic variation and significant pathological deficits and within different conditions based on the available data. Third, by providing real-time, dynamic visualizations and automated significance testing, the application lowers the barrier to entry for evidence-based biomarker screening in busy clinical environments.

Future work will focus on integrating cross-linguistic norms to support the transdiagnostic assessment of multilingual populations. Ultimately, ALBA represents a step toward a more objective, reproducible, and computationally-informed approach to speech and language pathology.

6. Data Availability

The raw linguistic transcripts used to derive the normative benchmarks in ALBA are sourced from the *AphasiaBank* and other clinical repositories within the *TalkBank* system (MacWhinney et al., 2011). Access to these raw data is governed by the TalkBank clinical data-sharing agreement, which requires researcher registration to protect patient confidentiality. Access to the aggregated reference measures (means, standard deviations, and deciles) for all 290 linguistic biomarkers across healthy and clinical cohort in the supplementary material as a machine-readable csv file and a stable URL to the ALBA interface are provided in `https://openbrainai.com/measures` clinical use.

7. References

S. Bengio and G. Heigold. Word embeddings for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1053–1057. Export Date: 20 February 2015.

Lucy Bryant, Alison Ferguson, and Elizabeth Spencer. 2016. Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical linguistics phonetics*, 30(7):489–518.

Lucy Bryant, Alison Ferguson, Megan Valentine, and Elizabeth Spencer. 2018. Implementation of discourse analysis in aphasia: investigating the feasibility of a knowledge-to-action intervention. *Aphasiology*, 33(1):31–57.

Madeline Cruice, Nicola Botting, Jane Marshall, Mary Boyle, Deborah Hersh, Madeleine Pritchard, and Lucy Dipper. 2020. Uk speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International journal of language communication disorders*, 55(3):417–442.

K. T. Cunningham and K. L. Haley. 2020. Measuring lexical diversity for discourse analysis in

aphasia: Moving-average type-token ratio and word information measure. *J Speech Lang Hear Res*, 63(3):710–721.

Gerasimos Fergadiotis. 2011. *Modeling lexical diversity across language sampling and estimation techniques*. Arizona State University.

Markus Forsberg, Dana Dannélls, Lars Borin, and Aleksandrs Berdicevskis. 2025. Background: Språkbanken text. In *Sixty years of Swedish computational lexicography / Dana Dannélls, Kristian Blensenius and Lars Borin (eds.)*, page 161–173. De Gruyter, Berlin.

W. N. Francis and H. Kucera. 1982. *Frequency analysis of English usage*. Houghton-Mifflin Company.

K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon. 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60. Fraser, Kathleen C Meltzer, Jed A Graham, Naida L Leonard, Carol Hirst, Graeme Black, Sandra E Rochon, Elizabeth eng MOP-82744/Canadian Institutes of Health Research/Canada Research Support, Non-U.S. Gov't Italy Cortex. 2014 Jun;55:43-60. doi: 10.1016/j.cortex.2012.12.006. Epub 2012 Dec 21.

T. E. A. Hansen, J. Praestegaard, T. Tjornhoj-Thomsen, M. Andresen, and B. Norgaard. 2022. Dementia-friendliness in danish and international contexts: A critical discourse analysis. *Gerontologist*, 62(1):130–141. Hansen, Tania E A Praestegaard, Jeanette Tjornhoj-Thomsen, Tine Andresen, Mette Norgaard, Birgitte eng FF2-R69-A1566/Danish Occupational Therapists Association ALZ/Alzheimer's Association/ 2021/05/18 Gerontologist. 2022 Jan 14;62(1):130-141. doi: 10.1093/geront/gnab056.

Swathi Kiran, L. Meier Erin, and P. Johnson Jeffrey. 2019. Neuroplasticity in aphasia: A proposed framework of language recovery. *Journal of Speech, Language, and Hearing Research*, 62(11):3973–3985. Doi: $10.1044/2019_{JSLHR-L-RSNP-19-0054}$.

Alexandra König, Aharon Satt, Alexander Sorin, et al. 2018. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, 1(1):112–124.

Houjun Liu, Brian MacWhinney, Davida Fromm, and Alyssa Lanzi. 2023. Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*, 66(7):2421–2433.

Brian MacWhinney. The childes project: Tools for analyzing talk: Transcription format and programs, vol. 1, 3rd ed.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307. Doi: 10.1080/02687038.2011.589893.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Brian Paltridge. 2006. *Discourse analysis : an introduction*. Continuum, London. GBA670244 bnb Brian Paltridge. Continuum discourse Includes bibliographical references and index. Formerly CIP. Uk.

Brielle C. Stark, Lucy Bryant, Charalambos Themistocleous, Dirk-Bart den Ouden, and Angela C. Roberts. 2023. Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology*, 37(5):761–784. Doi: 10.1080/02687038.2022.2039372.

Brielle C. Stark and Julia Fukuyama. 2021. Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Language, Cognition and Neuroscience*, 36(5):562–585. Doi: 10.1080/23273798.2020.1862258.

Charalambos Themistocleous. 2024. Open brain ai and language assessment. *Frontiers in Human Neuroscience*, 18.

Charalambos Themistocleous, Bronte Ficek, Kimberly Webster, Dirk-Bart den Ouden, Argye E. Hillis, and Kyrana Tsapkini. 2021. Automatic subtyping of individuals with primary progressive aphasia. *Journal of Alzheimer's Disease*, 79:1185–1194.

Charalambos Themistocleous and Brielle C. Stark. 2026. Language biomarker screening using ai: a transdiagnostic approach to brain. *Scientific Reports*.

Charalambos Themistocleous, Kimberly Webster, Alexandros Afthinos, and Kyrana Tsapkini. 2020. Part of speech production in patients with primary progressive aphasia: An analysis based on natural language processing. *American Journal of Speech-Language Pathology*, pages 1–15.

Donna C. Tippett, Katelyn Surrao, Kyriaki Neophytou, Hana Kim, Jessica Gallegos, Charalambos Themistocleous, Brenda Rapp, Argye E. Hillis, and Kyrana Tsapkini. 2025. Written picture descriptions distinguish variants of primary progressive aphasia. *Journal of Alzheimer's Disease*, page 13872877251376381. Doi: 10.1177/13872877251376381.

Maria Varkanitsa, Swathi Kiran, and Klaus Willmes. 2023. Measures of language and communication. *GJ Boyle, Y. Stern, DJ Stein, BJ Sahakian, CJ Golden*, pages 121–144.

On Automatic Detection of Cognitive Decline

Fabio Tamburini
FICLIT-University of Bologna
via Zamboni, 32, Bologna, Italy
fabio.tamburini@unibo.it

Abstract

Cognitive decline refers to the gradual loss of thinking abilities, including memory, attention, reasoning, and problem-solving. It can be a normal part of ageing or a symptom of conditions like dementia or Alzheimer's disease when it significantly interferes with daily life. Early diagnosis is crucial, as timely intervention can slow progression and improve quality of life. Emerging approaches such as Digital Linguistic Biomarkers, subtle changes in speech and language patterns captured through digital tools, offer a promising, non-invasive way to detect early signs of cognitive decline before more obvious symptoms appear and perform massive population screening. In this position paper, we contend that the prevailing paradigm for the automatic detection of cognitive decline, primarily relying on classifiers that analyse subjects' linguistic productions at a single point in time, is not the most effective approach. Instead, we advocate for a paradigm shift toward longitudinal analyses that track linguistic patterns over decades. To support this perspective, we present an experiment in which we compile and analyse a long-term corpus of spontaneous speech productions from well-known individuals, enabling insights into cognitive changes across extended time spans.

Keywords: Cognitive Decline, Digital Linguistic Biomarkers, Longitudinal Analyses

1. Introduction

Cognitive decline, ranging from mild memory impairment to severe loss of independence, represents a growing public health challenge with profound personal, societal, and economic consequences. Alzheimer's disease (AD), the most common cause of dementia, lies at the centre of this crisis. In 2025, approximately 7.2 million Americans aged 65 and older are estimated to be living with Alzheimer's dementia, and projections suggest this number could rise to nearly 13.8 million by 2060 (Alzheimer's Association, 2025). These figures reflect only the clinically visible portion of the problem and underestimate the true scope of Alzheimer's-related cognitive decline.

Biomarker-based studies reveal a much broader spectrum of disease. Around 5 million older adults may already have Alzheimer's-related dementia detectable through objective brain changes, while an additional 5–7 million may have Mild Cognitive Impairment (MCI) attributable to Alzheimer's pathology. Together, this implies that 10–12 million older Americans could be experiencing Alzheimer's-related cognitive decline even before reaching the stage of diagnosable dementia. MCI itself is common, affecting an estimated 12–18% of people aged 60 or older, and carries a substantial risk of progression: roughly 10–15% of individuals with MCI develop dementia each year, with about one-third progressing within five years (Alzheimer's Association, 2025).

Beyond prevalence, the burden of cognitive decline is immense. In 2025, healthcare and long-term care costs associated with Alzheimer's and other dementias are projected to reach $384 billion. When the value of unpaid caregiving, estimated at over $413 billion, is included, the total societal cost becomes staggering. This burden extends well beyond economics, deeply affecting caregivers, families, and communities. These realities highlight the urgency of addressing cognitive decline not only as a medical issue but also as a major social and policy concern.

A critical challenge lies in shifting the focus from treating advanced dementia to identifying individuals at earlier stages, such as MCI or even preclinical Alzheimer's disease. Pathological brain changes related to AD can begin decades before symptoms become apparent, yet current diagnostic approaches rely heavily on clinical presentation and expensive, specialised biomarker tests. As a result, opportunities for early intervention are often missed. This diagnostic gap underscores the need for feasible, scalable, and cost-effective screening methods that can be deployed at the population level. Early identification would allow clinicians to implement lifestyle interventions, monitor disease progression, and potentially apply emerging disease-modifying therapies when they are most effective. Population-wide screening could also reduce disparities by reaching individuals who lack access to specialised neurological care.

Alzheimer's disease is best understood as a continuous process rather than a set of discrete stages. It begins with a preclinical phase characterised by silent biological changes, such as amyloid and protein tau accumulation, without obvious symptoms. During this phase, individuals may experience Subjective Cognitive Decline (SCD), reporting perceived worsening of memory despite normal performance on standard tests. The next stage is MCI,

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 41–54
12 May 2026.

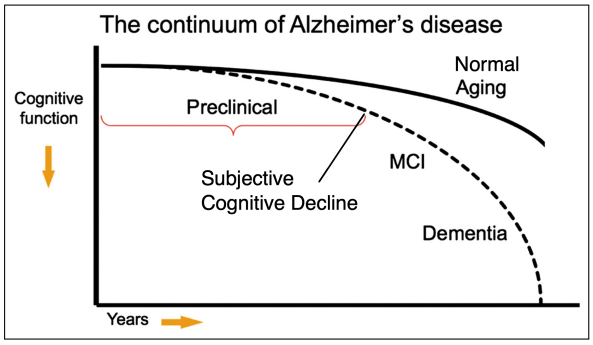


Figure 1: Model of the clinical trajectory of Alzheimer's disease (Picture adapted from Sperling et al. 2011).

often due to Alzheimer's pathology, where mild but noticeable cognitive symptoms emerge without significantly interfering with daily life. As the disease progresses, individuals move through stages of mild, moderate, and severe dementia, with increasing impairment in everyday functioning and eventual loss of independence. Importantly, cognitive decline unfolds gradually over many years, with disease-related trajectories showing steeper declines than normal ageing but remaining continuous in nature, as illustrated by Figure 1.

An important modifier of this trajectory is cognitive reserve (CR), a concept referring to the brain's resilience to age-related and pathological changes. Cognitive reserve helps explain why individuals with similar levels of brain pathology can show very different clinical outcomes. Factors such as education, intellectually demanding occupations, social engagement, and cognitively stimulating activities contribute to higher cognitive reserve, allowing some individuals to compensate more effectively and delay the onset of symptoms. Figure 2 illustrates the point clearly and, comparing this picture with Figure 1, the impact of cognitive reserve on diagnosis and disease progression should be evident.

Although cognitive reserve cannot be measured directly, it is inferred through proxies like educational attainment, occupational history, and lifestyle questionnaires (e.g. "CRIq" from Nucci et al. 2012). Higher cognitive reserve may delay diagnosis, even though underlying disease progression continues.

Within this context, novel screening approaches are gaining attention, particularly Digital Linguistic Biomarkers (DLBs) (Gagliardi et al., 2021). Language is a complex cognitive function supported by widespread brain networks (Catani et al., 2012; Hagoort, 2017; Hertrich et al., 2020), making it sensitive to subtle neural changes. DLBs consist of quantifiable linguistic and speech features, such as lexical diversity, syntactic complexity, fluency, semantic coherence, and acoustic properties, that can be automatically extracted using digital tools. Even mild disruptions in memory or executive function can manifest as detectable changes in everyday speech and writing.

Compared to traditional neuropsychological assessments, DLB-based methods are less resource-intensive and can be administered remotely and repeatedly at scale. Speech samples collected via smartphones or telehealth platforms can be analysed using natural language processing and machine learning techniques, enabling low-cost, ecologically valid monitoring of cognitive function over time. A growing body of evidence (see next section) shows that DLBs can reliably distinguish healthy ageing from MCI and early Alzheimer's disease, often detecting changes before overt behavioural symptoms appear and it has gained traction among researchers and clinicians as a means of obtaining fast, replicable, and objective proxy measures of mental disorders (Gagliardi, 2024).

Overall, the escalating burden of cognitive decline demands a paradigm shift toward early, population-level detection. By integrating scalable screening tools such as digital linguistic biomarkers with existing clinical approaches, healthcare systems may better address the personal, societal, and economic costs of Alzheimer's disease and related dementias.

1.1. State-of-the-art on the Automatic Detection of Cognitive Decline

In recent decades, advanced NLP techniques have been increasingly applied to the analysis of written texts, clinically elicited utterances, and spontaneous speech, with the aim of identifying DLBs of psychiatric and neurological disorders and automatically extracting linguistic features for pathology recognition, classification, and characterisation.

Computational methods have already proven effective in detecting linguistic indicators of cerebral functional disorders, including language alterations and disruptions linked to depression (Jiang et al., 2017; Stasak et al., 2019), focal brain lesions (Fergadiotis and Wright, 2011), Parkinson's disease (Benba et al., 2016; Sztahó and Vicsi, 2016; Arias-Vergara et al., 2018; Upadhya et al., 2019; Wang et al., 2022; Xue et al., 2023; Singh and Tripathi, 2024; Anap et al., 2025) and schizophrenia (Nenchev et al., 2024). They have also been successfully employed to detect prodromal dementia (MCI) (Roark et al., 2007, 2011; Satt et al., 2013; Vincze et al., 2016; dos Santos et al., 2017; Matsuda Toledo et al., 2018; Meilán et al., 2018; Tóth et al., 2018; Wang et al., 2019; Meilán et al., 2020; Wang et al., 2021; Gosztolya et al., 2021; Calzà et al., 2021; Ivanova et al., 2022; Egas-López et al., 2022; Moret-Tatay et al., 2023; Yamada et al., 2023;

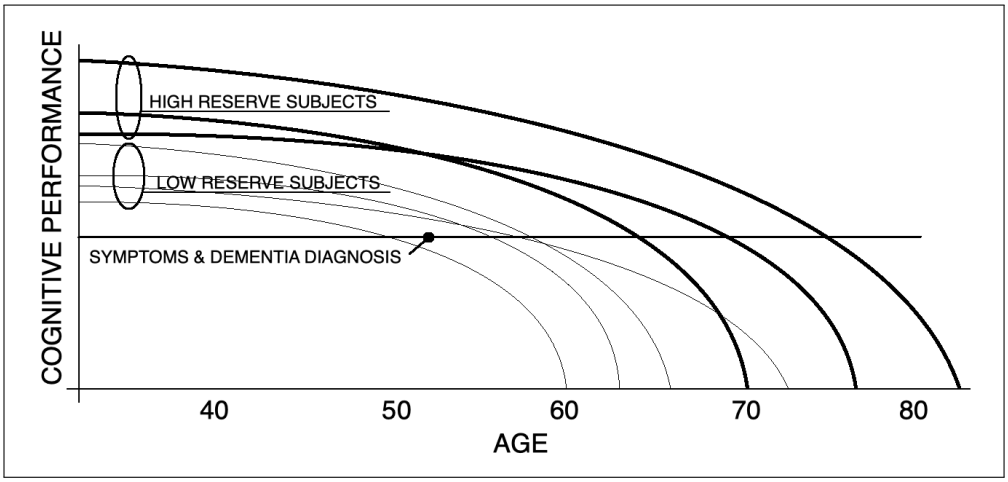


Figure 2: The graphs summarise evidence from observational studies indicating that individuals with higher and lower levels of education tend to differ in cognitive ability in early adulthood and even if, on average, show only small differences in rates of cognitive decline over time, the threshold of symptoms is reached at different ages (Picture adapted from Lövdén et al. 2020).

Favaro et al., 2024; Pourramezan Fard et al., 2024), as well as specific associated pathologies such as Alzheimer's disease (Jarrold et al., 2014; Fraser et al., 2016; Chinaei et al., 2017; López-de-Ipiña et al., 2015; Yancheva and Rudzicz, 2016; Sirts et al., 2017; Eyigoz et al., 2020; Yang et al., 2024; Huang et al., 2024; Kumar et al., 2025; Li et al., 2025), Primary Progressive Aphasia (PPA) (Fraser et al., 2014), and Frontotemporal Dementia (Jarrold et al., 2014; Coppieters et al., 2024).

Recent reviews (de la Fuente Garcia et al., 2020; Pulido et al., 2020; Petti et al., 2020; Ding et al., 2024; Cornacchia et al., 2025; Shankar et al., 2025; Shakeri and Farmanbar, 2025) outlines that, whereas neuropsychological tests and structured assessments often affect the naturalness of a subject's responses, the analysis of spontaneous spoken language offers an ecological and cost-efficient way to detect linguistic alterations in potential patients even within primary care settings.

Considering the cited literature, two main aspects emerge in addressing the problem:

- **Features/DLBs identification**: Many studies focus on defining or proposing a set of linguistic features capable of distinguishing potentially pathological subjects from healthy controls. Some studies manually extract features from speech or written samples, while others develop NLP systems to automate this process. Nearly all works employ statistical significance tests to identify the most promising linguistic indicators associated with the pathology.

- **Automatic classification**: Once relevant features are identified, several studies aim to construct automatic systems for pathology detection. Common machine/deep learning methods are employed, achieving varying levels of performance.

1.2. Past Shared Challenges

"ADReSS/ADReSSo" Challenges targets three difficult automatic prediction problems of societal and medical relevance, namely: detection of Alzheimer's Dementia, inference of cognitive testing scores, and prediction of cognitive decline by providing new and richly annotated English datasets to evaluate automatic systems (Luz et al., 2021b,a).

The recent "Prediction and Recognition Of Cognitive declinE through Spontaneous Speech" (PROCESS)[1] Signal Processing Grand Challenge at ICASSP-2025, proposes signal processing and prediction tasks to detect dementia via speech processing (Tao et al., 2025).

Both challenges introduced new datasets in which subjects' speech samples were classified into two or three classes following clinical judgments. Participants should apply their systems for classifying each sample/subject into one of the proposed classes. The best systems participating to these challenges obtained F1 classification results in the range from about 70% to 80%.

In general, the very large set of works published in the last years and devoted to the automatic detection of cognitive decline (see the survey papers cited before) present similar performance results: when dealing with the binary classification of healthy controls w.r.t. AD subjects, performance often are higher than 90% of correct classification, while, on the most challenging and definitely most

[1] https://processchallenge.github.io/

interesting problem of distinguishing MCI subjects from controls, it drops to around 75/80%.

As previously noted, it is crucial to detect the disease at its earliest stages, ideally when individuals present with MCI, or even earlier, when they experience only subjective and temporary memory difficulties. To be truly effective, the technological approaches described above must be capable of supporting large-scale screening across broad segments of the population. Unfortunately, current state-of-the-art systems do not yet provide sufficient reliability in identifying these early stages of dementia. In our view, this limitation is due more to challenges related to cognitive reserve than to the optimal combination of DLBs or classifier choice. Moreover, when analysing speech productions via DLB extraction to characterise individual speech profiles, subjects' speaking styles and specific accents, including those influenced by immigration from other countries, play a substantial role and can significantly blur the analysis of speech/language features.

These factors lead to considerable overlap between the two or three relevant classes in the feature space, resulting in unsatisfactory performance and limiting the applicability of such approaches for large-scale screening.

2. A Different Perspective

Building on the considerations outlined above, we argue that achieving the ultimate goal requires a true paradigm shift, in the Kuhnian sense. The cognitive reserve of an individual is extremely difficult to measure, as it is shaped by the entirety of his/her life experiences. Attempting to aggregate data from different subjects, even when they are classified within the same group, whether pathological (MCI/AD) or non-pathological (HC), creates challenges that machine learning classifiers cannot easily resolve with sufficient accuracy to enable large-scale population screening.

We believe a shift in perspective is needed: rather than designing systems that classify individuals "synchronically", at a single point in time, we should consider a "diachronic approach", examining each subject across the course of ageing. An ideal method would involve recording spontaneous speech samples at regular intervals, for example every two years after the age of 50, calculating the DLBs for each session, and assessing whether the individual shows signs of cognitive decline by comparing his/her current productions with his/her own past recordings.

This line of inquiry is not entirely new, as a limited number of studies have attempted to investigate cognitive decline using longitudinal data. For example, Laguarta and Subirana (2021) acknowledged the importance of longitudinal analyses and proposed a complex set of multimodal biomarkers that could, in principle, support such an approach. However, they did not present a longitudinal experiment due to the lack of suitable datasets. Petti et al. (2023) explored a simpler strategy, employing DLBs derived solely from written language (speech transcriptions), and demonstrated the promise of a longitudinal perspective. Gkoumas et al. (2024) introduced a multimodal longitudinal corpus spanning 12 months and conducted a DLB study on it, though the time span was too limited for significant changes to be observed. Comparable observations apply to the studies by Robin et al. (2023); Luz et al. (2021a), which are highly engaging but limited to a relatively short duration of 18 or 24 months. In another study, Petti and Korhonen (2024) created a novel longitudinal corpus by collecting interviews of famous individuals from YouTube, applying the same type of DLB analysis used in Petti et al. (2023). Despite the simplicity of their approach, the corpus itself is highly relevant to our objectives and will be examined in more detail in the following section. Finally, Chang et al. (2025) conducted an in-depth study on applying DLBs to track longitudinal trends, successfully distinguishing different trajectories between healthy controls and subjects with MCI. The main limitation, however, was the restricted number of longitudinal points, with data collected from only two visits/interviews.

The studies reviewed provide valuable groundwork for our perspective, highlighting the importance of longitudinal analyses in detecting cognitive decline. However, they all lack in adopting a fully subject-centred approach, which lies at the core of our proposal.

3. An Experiment to Support our View

To support our perspective, we designed an experiment that, given the nature of the available data, can only be regarded as a pilot study. The core of our proposal focuses on collecting subject data across the ageing process over an extended period of time, beginning, for instance, from the age of 50 onwards.

Unfortunately, no dataset currently exists that covers such an extended time span available for research purposes. There is, however, a notable exception: Petti and Korhonen (2024) introduced a longitudinal corpus - LoSST-AD - spanning a substantial portion of the subjects' lifespans, which would be ideal for our study. They compiled this resource by downloading interviews and other recordings from YouTube for ten well-known English-speaking individuals who had passed away from Alzheimer's disease, along with a matched set of healthy controls selected to reflect similar socio-

demographic profiles. However, for ethical reasons, the authors chose not to distribute the recordings themselves, making this valuable corpus only partially accessible (they released anonymised transcriptions only).

3.1. The μCLSD Dataset

Given the absence of an appropriate dataset to test our research hypothesis, we were compelled to construct a new linguistic resource, drawing inspiration from the work of Petti and Korhonen (2024). In the absence of large-scale longitudinal projects tracking subjects over the last 20–30 years of their lives, the only feasible approach is to rely on publicly available recordings of well-known individuals. Interviews with actors, writers, politicians, and other public figures represent a valuable source of material, potentially spanning decades and thus enabling extensive longitudinal analyses of speech production across the final stages of life.

We created the "Micro Corpus for the Longitudinal Study of Dementia" - μCLSD - by selecting 16 subjects, 8 who died from Alzheimer's disease (the AD group) and 8 who passed away due to other causes (the HC group), such as old age or illnesses not directly associated with cognitive impairment[2]. The sample is gender-balanced, and each AD subject is paired with an HC counterpart of the same gender and with a comparable professional background. Each group was further subdivided into four subjects speaking British English and four subjects speaking American English, in order to evaluate the approach across different varieties of English.

The selection of subjects was also guided by the availability of interviews on YouTube covering a wide time span of their lives, allowing us to reasonably assume that the earliest recordings were produced during periods unaffected by any cognitive disease.

From a technical perspective, we manually extracted audio fragments from these interviews, each lasting between one and one minute fifteen seconds. The interviewer's voice and external noise (e.g., music or applause) were removed to approximate the conditions of a spontaneous monologue in a quiet environment. All audio files were then resampled at 16 kHz, 16 bits and reduced to a single channel.

3.2. Our Pipeline for Extracting DLBs

Natural Language Processing (NLP) techniques and tools are playing an increasingly vital role in the medical field (Wang et al., 2020), supporting a wide spectrum of applications such as patient care, diagnostics, clinical coding, and patient-oriented services (Locke et al., 2021). In particular, there is a rising interest in leveraging automated speech and language analysis as a promising early indicator of pathological processes.

A newly built DLB pipeline (v2.0), based on our previous work (Gagliardi and Tamburini, 2022), processes audio signals to generate DLBs for each sample. It consists of two phases: preprocessing and feature extraction.

During the preprocessing phase, the input speech audio is transcribed relying on OpenAI Whisper-v3 using the "medium-en" model (Radford et al., 2023), then voice activity detection (Bredin et al., 2020), voiced segment identification, vowel-consonant distinction (Li et al., 2020), dependency parsing with UDPipe (Kondratyuk and Straka, 2019) and constituency parsing using STANZA (Qi et al., 2020) are performed on the input speech or its automatic transcription.

The feature extraction phase computes DLBs listed in Table 1 (please, refer to Calzà et al. 2021 for a detailed description) using the information obtained during preprocessing. These DLBs can be categorised into five groups: Acoustic, Rhythmic, Lexical, LIWC based counts (Pennebaker et al., 2015), and Syntactic DLBs, which, taken together, offer a fine-grained representation of the linguistic patterns related to subject cognitive abilities (Gagliardi and Tamburini, 2022).

Figure 3 depicts the overall structure of our pipeline. The tool can process three different kinds of inputs: spoken recordings (as a WAV audio file), raw written texts (TXT) transcriptions, or preprocessed texts in the CoNLL-U format containing morphosyntactic and syntactic analyses. Given a specific input type, either a WAV, TXT, or CoNLL file, the pipeline computes all the DLBs that can be derived from it. The larger set is obtained by providing the speech recording, alone or with the manual transcription (to bypass any mistake produced by the ASR module).

It is relevant to underscore that, for the experiments presented in this paper, we provide only the speech audio (WAV) file to the pipeline, thus any further computation must start from this single information and no manual effort is needed to process interview recordings for feature extraction. We conducted a series of tests to evaluate the effectiveness of the OpenAI Whisper-v3 model in transcribing English utterances, using data from the PROCESS Challenge as a benchmark (see Section 1.2). The model achieved a Word Error Rate (WER) of 8.7%, which appears satisfactory given that the speech is spontaneous and may include pathological traits. Our analysis showed that the primary

[2] Of course, in the absence of other clinical information, the HC group could, in principle, include individuals with undiagnosed cognitive impairment of some kind.

Acoustic DLBs (SPE)
Silence segments duration (M, MD, SD)
Speech segments duration (M, MD, SD)
Temporal regularity of voiced segments
Verbal Rate
Transformed Phonation Rate
Standardised Phonation Time
Standardised Pause Rate
Root Mean Square energy (M, SD)
Pitch (M, SD)
Spectral Centroid (M, SD)
Higuchi Fractal Dimension (M, SD)
Rhythmic DLBs (RHY)
Percentage of vocalic intervals - %V
SD of vocalic, ΔV, and cons., ΔC, interval durations
Pairwise Variability Index, raw, rPVI, and norm., nPVI
Variation coefficient for ΔV and ΔC
Lexical DLBs (LEX)
Content Density
Part-of-Speech rate
Reference Rate to Reality
Personal, Spatial and Temporal Deixis rate
Relative pronouns and negative adverbs rate
Lexical Richness: TTR, Brunet's and Honoré's Indexes
Action Verbs rate
Frequency-of-use tagging
Propositional Idea Density
Mean Number of words in utterances
Linguistic Inquiry and Word Count DLBs (LWC)
Language Metrics (e.g., words per sentence
Function Words (e.g., pronouns, articles, ...)
Affect Words (e.g., positive/negative emotion)
Cognitive Processes (e.g., insight, certainty, ...)
Perceptual processes (e.g., seeing, hearing, feeling)
Biological processes (e.g., body, health/illness, ...)
Personal concerns (e.g., work, leisure, money, ...)
Social Words (e.g., family, friends)
Punctuation (e.g., periods, commas, colons, ...)
Syntactic DLBs (SYN)
Number of dependent elements of the nouns (M, SD)
Global Dependency Distance (M, SD)
Syntactic complexity
Syntactic embeddedness: maximum tree depth (M, SD)
Utterance length (M, SD)

Table 1: The list of Digital Linguistic Biomarkers extracted by the pipeline. Some of these features are computed as means (M), medians (MD), and standard deviations (SD). Please refer to Calzà et al. (2021) for extended descriptions and computation details.

source of errors stems from the hyper-normalization tendency of ASR systems, which often remove or correct disfluencies, restarts, and repeated words to produce cleaner transcriptions. While this could pose a significant issue if disfluency counts were used as input features, we deliberately chose not to rely on such information. As a result, our system is only minimally affected by this limitation.

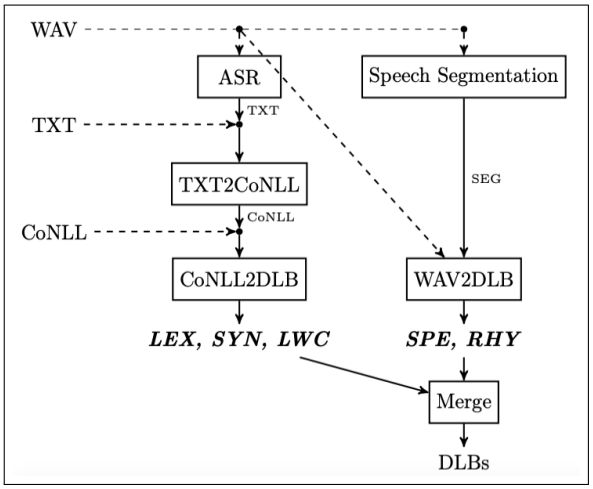


Figure 3: The whole structure of the Pipeline. Inputs can be provided as WAV, raw text, or CoNLL files. The modules are described in detail in Calzà et al. (2021) and Gagliardi and Tamburini (2022). In the experiments presented in this paper, we use only the audio WAV signal; all computations required to extract DLBs are carried out automatically.

3.3. Feature Processing & Selection

The Explainable Boosting Classifier (EBC)[3] (Lou et al., 2012) is an interpretable machine learning model from the family of Generalised Additive Models (GAMs). It combines the predictive power of boosting with the transparency of GAMs. Instead of learning a single complex function, it learns feature shape functions (one per feature, plus interactions, if allowed), which describe how each feature contributes to the prediction.

For feature ranking, the EBC provides global "importance scores" by measuring how much each feature contributes to the model's predictions across the dataset. This is typically done by (a) evaluating the magnitude of each feature's shape function (larger deviations indicate stronger impact) and (b) comparing across features to rank them by influence on the target outcome. This makes EBC especially useful in domains where both accuracy and interpretability matter, since it produces rankings along with human-readable explanations of how features affect predictions.

To determine the most relevant features for this task, we relied on the dataset provided for the PROCESS Challenge. The pipeline system described in the previous section took part in the competition and achieved strong performance, ranking first in the three-way classification of HC vs. MCI vs. AD (Zhang et al., 2025). Using the training and validation sets provided by this challenge, we carried out the following steps:

[3] https://interpret.ml

- extracted all DLBs listed in Table 1 using our software pipeline;
- normalised each feature using z-scores;
- applied the previously described EBC algorithm, retaining only DLBs with an importance score ≥ 0.01, resulting in the selection of 109 out of 126 features.

3.4. Drawing Cognitive Decline Profiles

Novelty detection with Local Outlier Factor (LOF) is a technique used to identify new data points that differ from the training distribution.

The LOF algorithm (Breunig et al., 2000) measures the local density deviation of a data point compared to its neighbours. The key idea is: points in dense regions are considered normal, while points in sparse regions, especially if their density is much lower than that of their neighbours, are considered novelties or outliers. For novelty detection (as opposed to outlier detection in training data), LOF is trained on "normal" examples only. New incoming samples are then scored: a score close to 1 means "normal", while larger scores indicate potential novelties/anomalies. This makes LOF useful in applications like fraud detection, fault monitoring, or detecting rare events in streaming data.

A short mathematical introduction for LOF in novelty detection could be described as:

- **k-distance and neighbors**. For each point x, find its k nearest neighbours $N_k(x)$ using distance $d(x, y)$. The *k-distance* is the distance to the k^{th} nearest neighbour.
- **Reachability distance**. For a point x and a neighbour y, the *reachability distance* is defined as:

$$\text{reach-dist}_k(x, y) = \max\big(d(x, y),\ k\text{-distance}(y)\big).$$

- **Local reachability density (LRD)**. The local reachability density of x is the inverse of the average reachability distance to its neighbours:

$$\text{lrd}_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} \text{reach-dist}_k(x, y)}.$$

- **LOF score**. The Local Outlier Factor compares the density of x with that of its neighbors:

$$\text{LOF}_k(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{\text{lrd}_k(y)}{\text{lrd}_k(x)}.$$

$\text{LOF}_k(x)$ can be easily interpreted as:

- $\text{LOF}_k(x) \approx 1$: x has similar density to neighbors $\Rightarrow$ x is in line with the training set (normal condition).
- $\text{LOF}_k(x) > 1$: x has much lower density $\Rightarrow$ x is an outlier or is different from points in the training set (novelty, thus a potentially pathologic condition).

We employed a LOF-based novelty detection approach to visualise the temporal evolution of subjects' cognitive abilities. Features were first normalised using z-scores, and the resulting values were then smoothed by computing a weighted moving average with a window of three samples, where weights reflected the temporal distance between sample pairs, to reduce the influence of transient spikes on neighbourhood contributions.

For each subject, the first four recordings were used as the training set for the LOF algorithm[4] assuming that these early samples represent speech unaffected by disease and thus serve as a reference for that individual. Subsequent recordings from the same subject were then compared against this reference LOF model, producing a score that reflects the degree of deviation from the reference. These scores were arranged to generate plots depicting each subject's cognitive functions trajectory over time in a way similar to Figure 1.

4. Results and Discussion

Figure 4 presents the computed cognitive function profiles. Comparing the first eight profiles of subjects diagnosed with AD to the eight profiles of healthy controls (HC) reveals notable differences: cognitive function profiles of AD subjects show significant deviations from the reference samples (the first four points in each profile) well before the corresponding markers of the official diagnosis. In contrast, the profiles of HC subjects remain stable until very advanced ages, reflecting the typical pattern of cognitive decline associated with normal ageing. The extracted DLBs and our proposed method for tracing their evolution over time appear to be sensitive to the differing cognitive trajectories of the two subject groups, allowing for precise detection of subtle speech variations linked to cognitive decline.

Our approach departs fundamentally from previous studies in the literature: rather than applying a classification process to determine a subject's cognitive status at a specific point in time comparing its DLBs with other subjects, we generate a continuous cognitive function profile that evolves across the ageing process considering only the DLBs of a single subject across time. This allows us to simulate an individual's cognitive trajectory and identify

[4] We relied on the Scikit-Learn LOF module. (https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_novelty_detection.html).

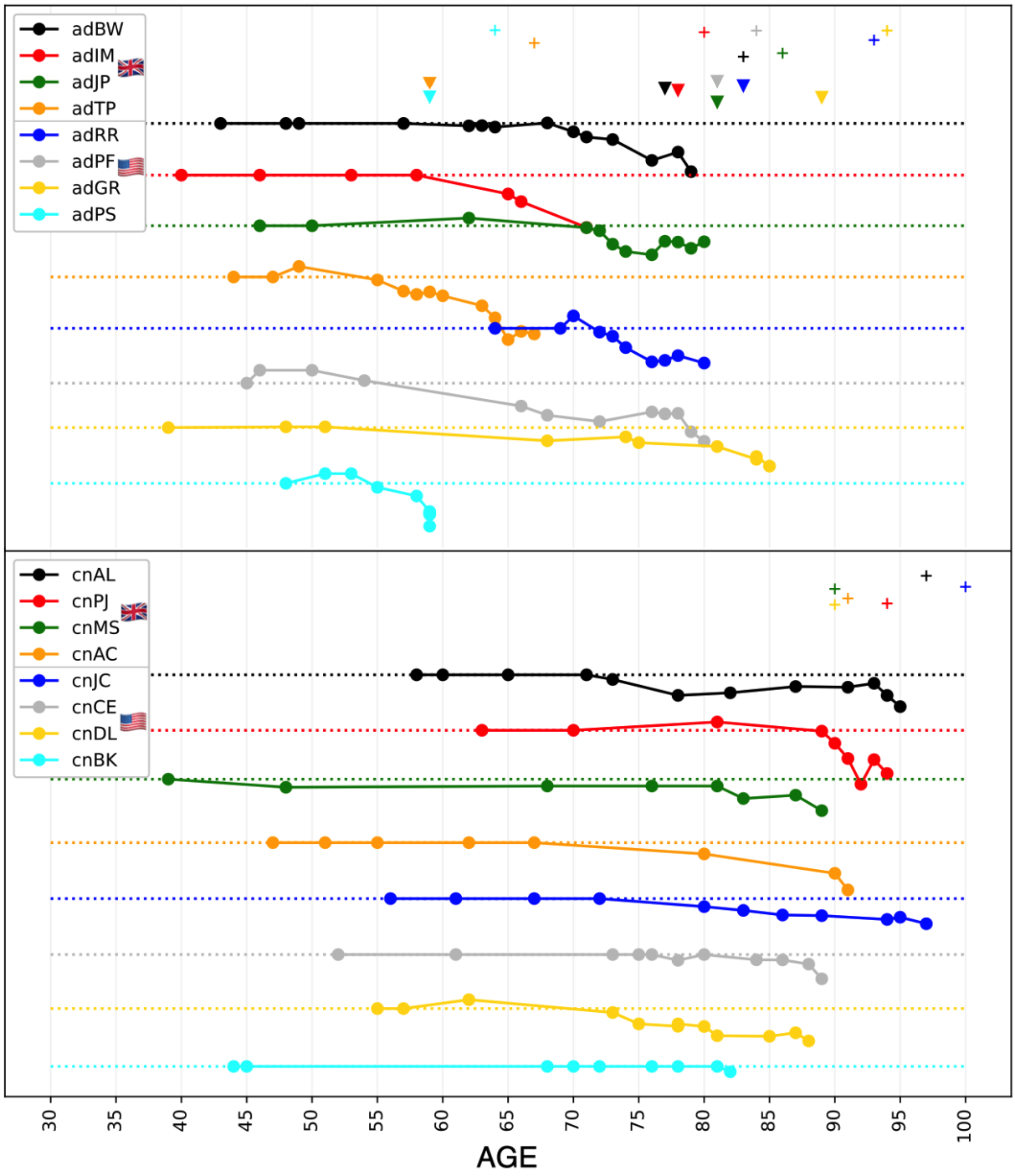


Figure 4: Evolution of cognitive functions over time for the sixteen subjects forming the μCLSD corpus. On top of each picture crosses mark the age of death, while triangles the year of first diagnosis for AD subjects. The top picture presents the profiles for the AD subjects (marked with 'ad' before name initials) while the bottom picture for HC (marked with 'cn'). Flags contained into the legenda indicate the English varieties spoken by the corresponding subjects.

significant deviations from his/her personal reference baseline, thereby detecting alterations relative to his/her normal cognitive status.

We do not envision this method being applied within specialised clinical practice. Instead, we propose it as a pilot approach to help define protocols for large-scale screening of ageing populations, aimed at detecting the earliest signs of cognitive decline. In this framework, general practitioners could use the method as a simple first-level tool and, when needed, refer individuals for specialist-administered neuropsychological assessments. Indeed, observed changes in a subject's cognitive profile do not necessarily indicate a true impair-

ment; rather, they may simply highlight a condition that warrants further evaluation by an expert to clarify the nature of these variations in cognitive functioning.

We do not claim that the proposed method can reliably screen entire populations or accurately distinguish individuals with cognitive impairment from healthy subjects. Rather, its purpose could be to support general practitioners in identifying potential concerns and referring these individuals to more specific and reliable assessments. In the absence of widely applicable large-scale screening tools, this approach could represent a practical solution, enabling general practitioners to serve as an initial filter for detecting individuals who may be at risk of cognitive impairment, even before any symptoms become apparent. For example, the method may capture within-person changes in speech, such as those related to general health variations or voice alterations due to other conditions, rather than patterns specifically associated with Alzheimer's disease. However, a general practitioner, being familiar with the individual's overall health status, may judge these signals as non-relevant and decide not to refer the person for further evaluation.

In the near future, we plan to evaluate whether the cognitive profile extraction method described in this paper remains an effective detection tool across different varieties of English (e.g. Australian English) and for other typologically distinct languages.

5. Limitations and Ethical Considerations

This pilot study is primarily intended to support the paradigm shift we propose for the early identification of cognitive decline. While the μCSLD dataset we collected is too limited in size and restricted to a single language to allow for broad generalisation, we contend that our approach provides a solid basis for devising new large-scale population screening methods based on DLBs, thereby addressing the limitations of the classification-based methods currently prevalent in the literature, and it favours the development of methods based on longitudinal analyses.

Regarding the corpus, subjects' data were anonymised in this paper but not in the dataset, as all recordings were obtained from publicly accessible sources on the Internet, primarily Wikipedia and YouTube. While voices may be recognisable and interview topics could potentially reveal identity, making full anonymisation of the dataset impossible, we believe that sharing such data is crucial to enable further research in this direction. Unfortunately, given that we selected dead subjects, it was not possible to collect any kind of consent from them.

For these reasons, the dataset will be available only upon request. It will include the audio recordings of the interviews, along with all references to the original sources (primarily URLs), and a clear listing of the subjects' names.

6. Acknowledgements

This study was funded by the European Union–NextGenerationEU programme through the Italian National Recovery and Resilience Plan– NRRP (Mission 4–Education and research), as a part of the project ReMind: an ecological, costeffective AI platform for early detection of prodromal stages of cognitive impairment (PRIN 2022, 2022YKJ8FP– CUP J53D23008380006).

7. Bibliographical References

Alzheimer's Association. 2025. 2025 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 21(4):e70235.

Sachin Anap, Satish Jondhale, Balasaheb Agarkar, and Sachin Chaudhari. 2025. Parkinson's disease detection from speech using combination of empirical wavelet transform and Hilbert transform. *International Journal of Speech Technology*, 28(1):185–194.

Tomas Arias-Vergara, Juan Camilo Vasquez Correa, Juan Rafael Orozco-Arroyave, and Elmar Nöth. 2018. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. *Speech Communication*, 101(101):11–25.

Achraf Benba, Abdelilah Jilbab, and Ahmed Hammouch. 2016. Voice Assessments for Detecting Patients with Parkinson's Diseases Using PCA and NPCA. *International Journal of Speech Technology*, 19(4):743–754.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, et al. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *Proc. ICASSP*, pages 7124–7128.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, page 93–104, New York, NY, USA. Association for Computing Machinery.

L. Calzà, G. Gagliardi, Rema Rossini Favretti, and F. Tamburini. 2021. Linguistic features and automatic classifiers for identifying Mild Cognitive Impairment and dementia. *Computer Speech & Language*, 65:101–113.

Marco Catani, Flavio Dell'Acqua, Alberto Bizzi, Stephanie J. Forkel, Steve C. Williams, Andrew Simmons, Declan G. Murphy, and Michel Thiebaut de Schotten. 2012. Beyond cortical localization in clinico-anatomical correlation. *Cortex*, 48(10):1262–1287.

Ho-Ling Chang, Thiri Wai, Yu-Shan Liao, Sheng-Ya Lin, Yu-Ling Chang, and Li-Chen Fu. 2025. A dual-modal fusion framework for detection of mild cognitive impairment based on autobiographical memory. *IEEE Journal of Biomedical and Health Informatics*, 29(6):4474–4485.

Hamidreza Chinaei, Leila Chan Currie, Andrew Danks, Hubert Lin, Tejas Mehta, and Frank Rudzicz. 2017. Identifying and Avoiding Confusion in Dialogue with People with Alzheimer's Disease. *Computational Linguistics*, 43(2):377–406.

Rosie Coppieters, Arabella Bouzigues, Lize Jiskoot, Maxime Montembeault, Boon Lead Tee, Jonathan D. Rohrer, Rose Bruffaerts, and et al. 2024. A systematic review of the quantitative markers of speech and language of the frontotemporal degeneration spectrum and their potential for cross-linguistic implementation. *Neuroscience & Biobehavioral Reviews*, 167:105909.

Ester Cornacchia, Aurora Bonvino, Giorgia Francesca Scaramuzzi, Daphne Gasparre, Roberta Simeoli, Davide Marocco, and Paolo Taurisano. 2025. Digital Screening for Early Identification of Cognitive Impairment: A Narrative Review. *WIREs Cognitive Science*, 16(4):e70009.

Sofia de la Fuente Garcia, Craig W. Ritchie, and Saturnino Luz. 2020. Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. 2024. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. *Artificial Intelligence Review*, 57:325.

Leandro B. dos Santos, Edilson Anselmo Corrêa Jr., Osvaldo N. Oliveira Jr, Diego R. Amancio, L. Mansur, and Sandra M. Aluísio. 2017. Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume Volume 1: Long Papers, pages 1284–1296. Association for Computational Linguistics.

José Vicente Egas-López, Réka Balogh, Nóra Imre, Ildikó Hoffmann, Martina Katalin Szabó, László Tóth, Magdolna Pákáski, János Kálmán, and Gábor Gosztolya. 2022. Automatic screening of mild cognitive impairment and Alzheimer's disease by means of posterior-thresholding hesitation representation. *Computer Speech & Language*, 75:101377.

Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, 28:100583.

Anna Favaro, Tianyu Cao, Najim Dehak, and Laureano Moro-Velázquez. 2024. Leveraging Universal Speech Representations for Detecting and Assessing the Severity of Mild Cognitive Impairment Across Languages. In *Proc. Interspeech 2024*, Kos, Greece.

Gerasimos Fergadiotis and Heather Harris Wright. 2011. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, 25:1414–1430.

Kathleen C. Fraser, Jed A. Meltzer, Naida L. Graham, Carol Leonard, Graeme Hirst, Sandra E. Black, and Elizabeth Rochon. 2014. Automated classification of Primary Progressive Aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, 49:407–422.

Gloria Gagliardi. 2024. Natural language processing techniques for studying language in pathological ageing: A scoping review. *Int J Lang Commun Disord*, 59:110–122.

Gloria Gagliardi, Dimitrios Kokkinakis, and Jon Andoni Duñabeitia. 2021. Editorial: Digital linguistic biomarkers: Beyond paper and pencil tests. *Frontiers in Psychology*, 12:752238.

Gloria Gagliardi and Fabio Tamburini. 2022. The automatic extraction of linguistic biomarkers as a viable solution for the early diagnosis of mental disorders. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5234–5242, Marseille, France. European Language Resources Association.

Dimitris Gkoumas, Bo Wang, Adam Tsakalidis, Maria Wolters, Matthew Purver, Arkaitz Zubiaga1, and Maria Liakata. 2024. A longitudinal multi-modal dataset for dementia monitoring and diagnosis. *Language Resources and Evaluation*, 58:883–902.

Gábor Gosztolya, Réka Balogh, Nóra Imre, José Vicente Egas-López, Ildikó Hoffmann, Veronika Vincze, László Tóth, Davangere P. Devanand, Magdolna Pákáski, and János Kálmán. 2021. Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech. *Computer Speech & Language*, 69:101215.

P. Hagoort. 2017. The core and beyond in the language-ready brain. *Neuroscience and biobehavioral reviews*, 81(Pt B):194–204.

Ingo Hertrich, Susanne Dietrich, and Hermann Ackermann. 2020. The margins of the language network in the brain. *Frontiers in Communication*, 5:93.

Lihe Huang, Hao Yang, Yiran Che, and Jingjing Yang. 2024. Automatic speech analysis for detecting cognitive decline of older adults. *Frontiers in Public Health*, 12.

Olga Ivanova, Juan José G. Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E. Llorente, and Nuria Carcavilla González. 2022. Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. *Computer Speech & Language*, 73:101341.

William L. Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided Diagnosis of Dementia Type through Computer-Based Analysis of Spontaneous Speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37. ACL - Association for Computational Linguistics.

Haihua Jiang, Bin Hu, Zhenyu Li, Lihua Yan, Tianyang Wang, Fei Liu, Huanyu Kang, and Xiaoyu Li. 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication*, 90:39–46.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.

M. Kumar, Sushant, and A. Yadav. 2025. Speech signal's phase information based Alzheimer's disease detection using deep learning. *International Journal of Speech Technology*, 28:397–410.

Jordi Laguarta and Brian Subirana. 2021. Longitudinal Speech Biomarkers for Automated Alzheimer's Detection. *Frontiers in Computer Science*, 3.

Chuyuan Li, Raymond Li, Thalia S. Field, and Giuseppe Carenini. 2025. Delta-KNN: Improving demonstration selection in in-context learning for Alzheimer's disease detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25807–25826, Vienna, Austria. Association for Computational Linguistics.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, et al. 2020. Universal phone recognition with a multilingual allophone system. In *Proc. ICASSP*, pages 8249–8253.

Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, and Gareth B. Kitchen. 2021. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, 38:4–9.

Karmele López-de-Ipiña, Jordi Solé-Casals, Harkaitz Eguiraun, J.B. Alonso, C.M. Travieso, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage, and Blanca Beitia. 2015. Feature selection for spontaneous speech analysis to aid in alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, 30(1):43 – 60.

Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 150–158, New York, NY, USA. Association for Computing Machinery.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021a. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Interspeech 2021*, pages 3780–3784.

Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. 2021b. Editorial: Alzheimer's Dementia Recognition through Spontaneous Speech. *Frontiers in Computer Science*, Volume 3 - 2021.

Martin Lövdén, Laura Fratiglioni, M. Maria Glymour, Ulman Lindenberger, and Elliot M. Tucker-Drob. 2020. Education and Cognitive Functioning Across the Life Span. *Psychological Science in the Public Interest*, 21(1):6–41.

Cintia Matsuda Toledo, Sandra Maria Aluisio, Leandro Borges dos Santos, Sonia Maria Dozzi Brucki, Eduardo Sturzeneker Trés, Maira Okada de Oliveira, and Letícia Lessa Mansur. 2018. Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:31–40.

J.J.G. Meilán, F. Martínez-Sánchez, J. Carro, N. Carcavilla, and O. Ivanova. 2018. Voice Markers of Lexical Access in Mild Cognitive Impairment and Alzheimer's Disease. *Current Alzheimer Research*, 15:111–119.

Juan J. G. Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E. Llorente, and Juan Carro. 2020. Changes in the Rhythm of Speech Difference between People with Nondegenerative Mild Cognitive Impairment and with Preclinical Dementia. *Behavioural Neurology*, 2020(1):4683573.

Carmen Moret-Tatay, Isabel Iborra-Marmolejo, María José Jorques-Infante, Gloria Bernabé-Valero, María José Beneyto-Arrojo, and Tatiana Quarti Irigaray. 2023. A pilot screening for cognitive impairment through voice technology (WAY2AGE). *BMC Psychology*, 11(1):170.

Ivan Nenchev, Tatjana Scheffler, Marie de la Fuente, Heiner Stuke, Benjamin Wilck, Sandra Anna Just, and Christiane Montag. 2024. Linguistic markers of schizophrenia: a case study of Robert Walser. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 41–60, St. Julians, Malta. Association for Computational Linguistics.

M. Nucci, D. Mapelli, and S. Mondini. 2012. The cognitive Reserve Questionnaire (CRIq): a new instrument for measuring the cognitive reserve. *Aging clinical and experimental research*, 24:218–226.

James Pennebaker, Roger Booth, Ryan Boyd, and Martha Francis. 2015. Linguistic inquiry and word count: Liwc2015. Technical report.

U. Petti, S. Baker, and A. Korhonen. 2020. A systematic literature review of automatic alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Ulla Petti, Simon Baker, Anna Korhonen, and Jessica Robin. 2023. The Generalizability of Longitudinal Changes in Speech Before Alzheimer's Disease Diagnosis. *Journal of Alzheimer's Disease*, 92(2):547–564.

Ulla Petti and Anna Korhonen. 2024. LoSST-AD: A longitudinal corpus for tracking Alzheimer's disease related changes in spontaneous speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10813–10821, Torino, Italia. ELRA and ICCL.

Ali Pourramezan Fard, Mohammad H. Mahoor, Muath Alsuhaibani, and Hiroko H. Dodge. 2024. Linguistic-based Mild Cognitive Impairment detection using Informative Loss. *Computers in Biology and Medicine*, 176:108606.

María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. 2020. Alzheimer's disease and automatic speech analysis: A review. *Expert Systems with Applications*, 150:113213.

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*. ACL, Stroudsburg (PA).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *PMLR*, pages 28492–28518.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting Mild Cognitive Impairment. In Kevin Bretonnel Cohen, Dina Demner-Fushman, Carol Frieman, Lynette Hirschman, and John Pestian, editors, *Proceedings of the Workshop BioNLP 2007: Biological, translational, and clinical language processing*, pages 1–8. ACL, Stroudsburg (PA).

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey A. Kaye. 2011. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE Transactions on Audio Speech, and Language Processing*, 19(7):2081–2090.

Jessica Robin, Mengdan Xu, Aparna Balagopalan, Jekaterina Novikova, Laura Kahn, Abdi Oday, Mohsen Hejrati, Somaye Hashemifar, Mohammadreza Negahdar, William Simpson, and Edmond Teng. 2023. Automated detection of progressive speech changes in early alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 15(2):e12445.

Aharon Satt, Alexander Sorin, Orith Toledo-Ronen, Oren Barkan, Ioannis Kompatsiaris, Athina Kokonozi, and Magda Tsolaki. 2013. Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia. In *Proceedings of Interspeech 2013*, pages 1692–1696. ISCA, Grenoble.

Arezo Shakeri and Mina Farmanbar. 2025. Natural language processing in alzheimer's disease research: Systematic review of methods, data, and efficacy. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 17(1):e70082.

Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. 2025. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, 3(2):100205.

Nutan Singh and Priyanka Tripathi. 2024. An ensemble technique to predict Parkinson's disease using machine learning algorithms. *Speech Communication*, 159:103067.

Kairit Sirts, Olivier Piguet, and Mark Johnson. 2017. Idea density for predicting Alzheimer's disease from transcribed speech.

Reisa A. Sperling, Paul S. Aisen, Laurel A. Beckett, David A. Bennett, Suzanne Craft, Anne M. Fan, Takeshi Iwatsubo, Clifford R. Jack Jr., Jeffrey Kaye, Thomas J. Montine, Denise C. Park, Eric M. Reiman, Christopher C. Rowe, Eric Siemers, Yaakov Stern, Kristine Yaffe, Maria C. Carrillo, Bill Thies, Marcelle Morrison-Bogorad, Molly V. Wagster, and Creighton H. Phelps. 2011. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):280–292.

Brian Stasak, Julien Epps, and Roland Goecke. 2019. Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis. *Speech Communication*, 115:1–14.

Dávid Sztahó and Klára Vicsi. 2016. Estimating the Severity of Parkinson's Disease Using Voiced Ratio and Nonlinear Parameters. In Pavel Král and Carlos Martín-Vide, editors, *Statistical Language and Speech Processing*, pages 96–107. Springer International Publishing, Cham.

Fuxiang Tao, Bahman Mirheidari, Madhurananda Pahar, Sophie Young, Yao Xiao, Hend Elghazaly, Fritz Peters, Caitlin Illingworth, Dorota Braun, Ronan O'Malley, Simon Bell, Daniel Blackburn, Fasih Haider, Saturnino Luz, and Heidi Christensen. 2025. Early Dementia Detection Using Multiple Spontaneous Speech Prompts: The PROCESS Challenge. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

László Tóth, Ildiko Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. 2018. A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. *Current Alzheimer Research*, 15:1–10.

Savitha S. Upadhya, A. N. Cheeran, and J. H. Nirmal. 2019. Discriminating parkinson diseased and healthy people using modified mfcc filter bank approach. *International Journal of Speech Technology*, 22(4):1021–1029.

Veronika Vincze, Gábor Gosztolya, László Tóth, Ildikó Hoffmann, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. 2016. Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 181–187. Association for Computational Linguistics.

Jing Wang, Huan Deng, Bangtao Liu, Anbin Hu, Jun Liang, Lingye Fan, Xu Zheng, Tong Wang, and Jianbo Lei. 2020. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on pubmed. *Journal of Medical Internet Research*, 22(1):e16816.

Meng Wang, Yanxia Wen, Shicong Mo, Liqiong Yang, Xiaqing Chen, Man Luo, Hongdian Yu, Fan Xu, and Xianwei Zou. 2022. Distinctive acoustic changes in speech in Parkinson's disease. *Computer Speech & Language*, 75:101384.

Tianqi Wang, Yin Hong, Quanyi Wang, Rongfeng Su, Manwa Lawrence Ng, Jun Xu, Lan Wang, and Nan Yan. 2021. Identification of Mild Cognitive Impairment Among Chinese Based on Multiple Spoken Tasks. *Journal of Alzheimer's Disease*, 82(1):185–204.

Tianqi Wang, Chongyuan Lian, Jingshen Pan, Quanlei Yan, Feiqi Zhu, Manwa L. Ng, Lan Wang, and Nan Yan. 2019. Towards the Speech Features of Mild Cognitive Impairment: Universal Evidence from Structured and Unstructured Connected Speech of Chinese. In *Proc. Interspeech 2019*, pages 3880–3884.

Zaifa Xue, Huibin Lu, Tao Zhang, Jiahui Xu, and Xiaonan Guo. 2023. A local dynamic feature

selection fusion method for voice diagnosis of parkinson's disease. *Computer Speech & Language*, 82:101536.

Yasunori Yamada, Kaoru Shinkawa, Miyuki Nemoto, Kiyotaka Nemoto, and Tetsuaki Arai. 2023. A mobile application using automatic speech analysis for classifying Alzheimer's disease and mild cognitive impairment. *Computer Speech & Language*, 81:101514.

Maria Yancheva and Frank Rudzicz. 2016. Vector-space topic models for detecting Alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2337–2346. Association for Computational Linguistics.

X. Yang, K. Hong, D. Zhang, and K. Wang. 2024. Early diagnosis of Alzheimer's Disease based on multi-attention mechanism. *PLoS ONE*, 19(9):e0310966.

Shibingfeng Zhang, Nadia Khlif, Marcello Ferro, Gloria Gagliardi, and Fabio Tamburini. 2025. Cognitive Decline Detection using DLB Extraction Pipelines. In *Proceedings of Prediction and Recognition Of Cognitive declinE through Spontaneous Speech (PROCESS) Signal Processing Grand Challenge, ICASSP-25*, Hyderabad, India. IEEE.

Benchmarking NLP-supported Language Sample Analysis for Swiss Children's Speech

Anja Ryser, Yingqiang Gao, Sarah Ebling*
Department of Computational Linguistics
University of Zurich, Switzerland
{ryser, yingqiang.gao, ebling}@cl.uzh.ch

Abstract

Language sample analysis (LSA) is a process that complements standardized psychometric tests for diagnosing, for example, developmental language disorder (DLD) in children. However, its labour-intensive nature has limited its use in speech-language pathology practice. We introduce an approach that leverages natural language processing (NLP) methods that do not rely on commercial large language models (LLMs) applied to transcribed speech data from 119 children in the German-speaking part of Switzerland with typical and atypical language development. This preliminary study aims to identify optimal practices that support speech-language pathologists in diagnosing DLD more efficiently with active involvement of human specialists. Preliminary findings underscore the potential of integrating locally deployed NLP methods into the process of semi-automatic LSA.

Keywords: language sample analysis, developmental language disorder, automatic speech recognition.

1. Introduction

Developmental language disorder (DLD) is a neurodevelopmental condition, commonly diagnosed in children, that significantly affects an individual's ability to acquire and use spoken and written language, despite typically developed intelligence and no obvious sensory or neurological impairments or inadequate language exposure (Tomblin et al., 1996; Bishop, 2006; Lüke et al., 2023; van Wijngaarden et al., 2024).

As a recommended part of DLD diagnosis, language sample analysis (LSA) aims at evaluating the spontaneous[1] language production skills of children (Gallagher and Hoover, 2020; Ramos, 2024). It involves collecting and analysing samples of language during conversation, storytelling, play, or other activities. LSA provides detailed information about a person's linguistic abilities, including vocabulary, grammar, sentence structure, and pragmatic language use. These insights can then be used in diagnostics, setting therapeutic goals and monitoring progress.

Despite being an effective tool for practice, LSA is not often used by speech-language pathologists due to its time- and effort-intensive process (Klatte et al., 2022; Bawayan et al., 2022). Modern NLP methods and machine learning can help to alleviate some of these challenges with their time efficient approaches to big amounts of data

In this study, we evaluate the zero-shot capability of non-commercial NLP methods, namely ASR and part-of-speech (POS) tagging on data collected from 110 children living in Switzerland. These NLP methods are the foundation on which analyses of the speech and language samples, such as measurements for lexical density and diversity and many others, are based on. To ensure the quality and reliability of these analyses, it is crucial to have trustworthy NLP methods with no commercial LLMs involved, as described in this study.

Transcription	Variant	WER	CER	MER	WIL
Original	Swiss German	81.0	80.0	49.9	94.8
	Swiss Std. German	48.7	47.8	36.1	59.5
Normalized	Swiss German	57.2	55.7	38.6	73.8
	Swiss Std. German	45.6	44.8	35.2	54.7

Table 1: Average ASR results on Whisper transcriptions for original and normalized text.

With data collected from 119 children living in Switzerland, we present a case study demonstrating that NLP approaches that do not rely on commercial LLMs can effectively assist speech-language pathologists in identifying critical linguistic patterns essential for LSA. Results from analyses of Swiss German and Swiss Standard German speech transcriptions highlight the potential of automating LSA. To achieve the high quality performance needed in a clinical setting, our results also show the need for specific training and fine-tuning.

Our **main contributions** are three-fold: 1) We demonstrate a possible annotation process in semi-automated LSA for the elicitation of clinical linguistic features of DLD in Swiss German; 2) We release the first case study dataset in Swiss German and Swiss Standard German containing six speech transcriptions together with their annotations; 3)

*Corresponding author. Dataset on SWISSUBase

[1]We acknowledge that elicited language is never completely "spontaneous"; nevertheless, the term is common in connection with LSA.

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 55–73
12 May 2026.

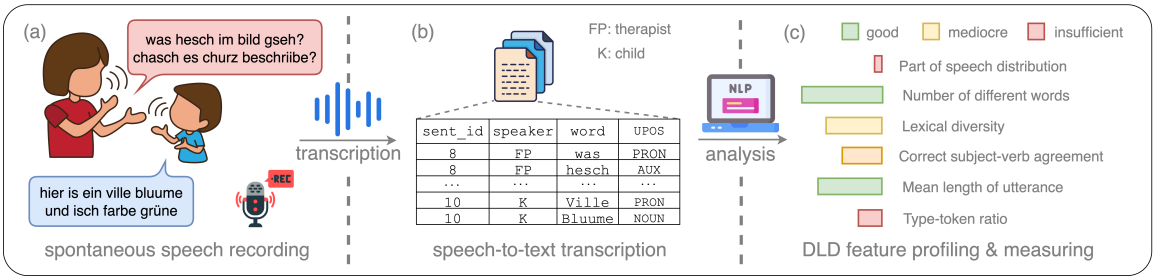


Figure 1: Our pipeline of LSA with NLP-supported approaches for diagnosis of DLD. **(a) Spontaneous speech recording**: a speech-language pathologist interacts with the child in a naturalistic setting and both of their speeches are recorded; **(b) Speech-to-text transcription**: the recordings are (automatically) converted into text and post-corrected by the speech-language pathologist and further (automatically) tokenized to words; **(c) DLD feature profiling and measuring**: approaches such as POS tagging, dependency parsing, stemming and lemmatization, etc., are applied to create the DLD feature profiles, where various linguistic measures are computed to evaluate the language abilities of children. The final diagnostic decision is made by the human expert (i.e., the speech-language pathologist), taking into account the output of the pipeline as well as other criteria. The speech utterances demonstrated are in Swiss German.

We provide an assessment of the effectiveness of NLP methods not based on commercial LLMs for LSA on both Swiss German and Swiss Standard German.

2. Related Work

2.1. Speech Transcription

Transcribing children's speech into text is a critical first step in LSA. However, advanced end-to-end approaches, such as ASR, still face substantial challenges due to the high inter- and intra-speaker variability in pronunciation, vocabulary, and speech rate among children across different ages as well as higher fundamental frequencies in children (Potamianos et al., 1997; Bhardwaj et al., 2022). Smith et al. (2017) trained a deep neural network on out-of-domain adult speech data, which was subsequently fine-tuned using speech data from children with DLD. Similarly, Rumberg et al. (2021) proposed a framework for age-invariant training, leveraging age-independent patterns derived from both adult and child speech. Jain et al. (2023) studied the adaption of the Whisper model (Radford et al., 2023) to child speech via self-supervised fine-tuning. Pokel et al. (2025) introduced a novel algorithm for ASR tailored to dysarthric German speech, which restructures word-level utterances into sentence-level sequences. This approach demonstrates promising results in improving the accessibility of speech transcriptions.

These approaches, often leveraging the power of transfer learning by utilizing adult speech data, have shown effectiveness in increasing performance on children's speech data. However, the fundamental lack of data collected from children with DLD still limits the applicability of ASR in LSA. In recent years, ASR systems for low-resource languages, such as Swiss German, have been developed (Kew et al., 2020; Nigmatulina et al., 2020; Arabskyy et al., 2021; Schraner et al., 2022; Timmel et al., 2024), supported by the availability of corpora of Swiss German speech data (Plüss et al., 2020; Dogan-Schönberger et al., 2021; Plüss et al., 2022, 2023; Stucki et al., 2025). However, due to licensing limitations, these models are not currently available and thus the ability of Swiss German models to generalize to speech data of children with DLD remains largely unexplored. Additionally, finetuning of commercial models running on commercial computation services with children's speech data is heavily restricted due to legal and ethical considerations and limitations. (Liu et al., 2024). As Whisper models perform well on Swiss German in a zero-shot setting (Dolev et al., 2024) and their availability, a Whisper model was used for this research for speech transcription.

2.2. Feature Analysis

Initial efforts to develop semi-automated LSA approaches have emerged in the past years, with a primary focus on English speech data. Gabani et al. (2011) analysed speech data collected from monolingual English-speaking children and proposed NLP methods to predict the presence of DLD. The study utilized eight categories of linguistic features (later expanded by Hassanali et al. (2012) to include syntactic and semantic features) to train language models as predictors, which provided important aspects for future research. Solorio (2013) provided a concise summary of the types of NLP-features employed in LSA for the diagno-

sis of DLD and highlighted questions for future research. Lüdtke et al. (2023) described the ideal hypothetical system capable of recording spontaneous speech of children while effectively separating background noise and speech from non-target individuals. This system can transcribe and segment recorded speech, offering a wide range of measurements, including environmental factors of recording, DLD profiles, and detailed analyses across various linguistic structures and elements.

These studies have laid the groundwork for advancing semi-automated LSA approaches, highlighting key linguistic features, methodological considerations, and future directions for improving DLD diagnosis. However, none of these works investigated LSA for children's speech in Swiss German.

DLD features can manifest in all linguistic categories. At the moment we are focusing on the grammatical level, which can be analysed based on transcripts of speech. As part of future research, phonetic and phonological features could also be analysed directly on the speech recordings.

In this work, we show that it is possible to perform LSA using NLP methods that are not based on commercial LLMs and are both ethical and effective. While we acknowledge LLMs are likely to dominate in LSA, we argue that other NLP methods still deliver good results without causing potential ethical issues.

3. Data

The dataset is scheduled to be published on SWISSUbase (DOI: https://doi.org/10.48656/rqf2-sq76) and will be released in the near future for public access.

3.1. Data Collection

Speech data collected from children are highly sensitive and require careful handling. In compliance with the regulations of the responsible research committee, we obtained the necessary ethical approval to collect speech utterances, accompanied by signed consent forms from the children's parents.[2] Speech utterances were recorded in both therapeutic and naturalistic settings (i.e., kindergarten), capturing spontaneous interactions between one therapist and one child. To ensure privacy, the data were stored in an anonymized format, with no metadata linked to identifiable codes.

We collected speech samples of duration between 10 and 20 minutes from 119 children with typical and atypical Swiss German and Swiss Standard German speech and 25 speech-language pathologists. Of these, we obtained permission to publish the data of 41 recordings. Although the dataset is relatively small, our objective is to initiate research into the application of semi-automated LSA for Swiss German speech utterances based on this dataset. The recordings were made with phones in standard quality and stored as mp3-files.

[2]For further details, please refer to the Ethics Statement section.

	Swiss German	Swiss Std. German
# recordings	91	19
# hours	17:57	3:35
# utterances	16,553	3,014
# words	126,733	21,356

Table 2: Statistics of the collected data. The table summarizes the total duration of recordings, the number of utterances and individual words for both speakers.

3.2. Data Statistics

Table 2 presents the overall statistics of the collected speech data. Speech recordings were obtained in both Swiss German and Swiss Standard German. The ages of the children range from four to eight, encompassing an important phase in the assessment of and intervention for DLD (Sansavini et al., 2021). The speech data were transcribed by students of speech and language therapy and subsequently verified by professionals native in both Swiss German and Swiss Standard German.

We engaged with children living across Switzerland. Therefore, the collected recordings are composed of different Swiss German dialects such as "*Züridüütsch*" (dialect spoken in the Canton of Zurich) and "*Baseldüütsch*" (dialect spoken in the Canton of Basel).

4. Methods

Starting from the collected recordings of spontaneous speech of Swiss children, we showcase our data processing methods specifically used for semi-automated LSA.

4.1. Speech Transcription

The first task is to transcribe speech into text based on the raw audio recordings. To do this, we apply two approaches:

Manual transcription by human experts. We recruited 13 students majoring in speech and language therapy, who had received training in the transcription method used for this study. For the transcriptions we use an adapted form of the *Dieth-Schreibung* (*Dieth spelling*) (Dieth, 1986) with added annotation of features important for

language and speech pathology, such as annotation of stress, unintelligible sequences, pauses, mazes, and overlaps in turn taking. Each transcript was created by one student and checked by another. The manual transcriptions served as the ground-truth reference for our study. See Table 20 in Appendix E for dataset examples. Additionally, the authors normalized the transcript manually to compare the influence on non-standard orthography and an assimilation to Standard German. Where possible, orthographically correct versions of words were used while keeping the word order intact.

Transcription using ASR models. We deployed a Whisper model (Radford et al. (2023), Hugging Face checkpoint `openai/whisper-small`[3]) **locally** to transcribe speech recordings into spoken sentences. This lightweight Whisper model, with 244 million parameters, was selected for its ease of deployment on local computers to prevent data leakage and its multilingual transcription capabilities and before the data was collected.

Automatically transcribing the speech samples used in this study posed three challenges: (1) Most large ASR systems are not trained with Swiss German data; (2) Most ASR systems underperform in transcribing speech from children (Bhardwaj et al., 2022); (3) The speech transcribed contains non-standard, atypical or wrong grammar due to the atypical language development of the children. Combining these three difficulties is challenging, requesting highly specifically trained models to solve the task reliably.

Manual transcription of spontaneous speech demands substantial knowledge and expertise in LSA and considerable time investment. Therefore, we aimed to evaluate the performance of the state-of-the-art Whisper model on the speech transcription task, given its potential to significantly reduce the workload of speech-language pathologists in practice.

4.2. Part-of-speech (POS) Tagging

Since POS tagging delivers information for certain linguistic features used for the feature analysis as described in Section 2.2, and thus helps in identifying morphosyntactic errors in language samples, our second task is to perform POS tagging for transcriptions in Swiss German and Swiss Standard German. For an overview over both POS tag sets used in this work, see Table 3.

For Swiss German transcriptions, we applied two BERT-based models (Aepli and Sennrich, 2022), trained with two different POS tag sets: 1) `swiss_german_pos_model`, trained with the Universal POS tags (UPOS[4]), and 2)

UPOS Tags	Feature	UPOS Tags	Feature
ADJ	adjective	ADP	adposition
ADV	adverb	AUX	auxiliary
CCONJ	co. conjunction	DET	determiner
INTJ	interjection	NOUN	noun
NUM	numeral	PART	particle
PRON	pronoun	PROPN	proper noun
PUNCT	punctuation	SCONJ	sub. conjunction
SYM	symbol	VERB	verb
X	other		

STTS Tags	Feature
ADJA	Attributive adjectives
ADJD	Predicative or adverbial adjectives
APPO	Postpositions
APPR	Prepositions
APPRART	Prepositions with an article
APZR	Circumpositions (right part)
ADV	True adverbs
ART	Definite/indefinite articles
CARD	Cardinal numbers
KOKOM	Comparative particles
KON	Coordinating conjunctions
KOUI	Subordinating conjunctions with infinitive
KOUS	Subordinating conjunctions with clauses
ITJ	Interjections
NE	Proper nouns
NN	Common nouns
PTKA	Particles with adjectives or adverbs
PTKANT	Response particles
PTKNEG	Negation particles
PTKVZ	Separable verb prefixes
PTKZU	“zu” before infinitives
PAV	Pronominal adverbs
PDAT	Attributive demonstrative pronouns
PDS	Substituting demonstrative pronouns
PIAT	Attributive indefinite pronouns without determiners
PIDAT	Attributive indefinite pronouns with determiners
PIS	Substituting indefinite pronouns
PPER	Non-reflexive personal pronouns
PPOSAT	Attributive possessive pronouns
PPOSS	Substituting possessive pronouns
PRF	Reflexive personal pronouns
PRELAT	Attributive relative pronouns
PRELS	Substituting relative pronouns
PWAT	Attributive interrogative pronouns
PWAV	Adverbial interrogative pronouns
PWS	Substituting interrogative pronouns
TRUNC	Truncation
FM	Foreign language material
XY	Non-words
VAFIN	Auxiliary finite verbs
VMFIN	Modal finite verbs
VVFIN	Full finite verbs
VAIMP	Auxiliary imperative verbs
VVIMP	Full imperative verbs
VAINF	Auxiliary infinitives
VMINF	Modal infinitives (substitute infinitive)
VVINF	Full infinitives
VVIZU	Infinitives with “zu”
VAPP	Auxiliary past participles
VMPP	Modal past participles
VVPP	Non-inflected full past participles

Table 3: Overview of two POS tagging systems for German, UPOS (top) and STTS (bottom). The core difference is that STTS contains the level of detail required for LSA.

[3]Model available at https://huggingface.co/openai/whisper-small

[4]https://universaldependencies.org/u/pos/

swiss_german_stts_pos_model, trained with the Stuttgart-Tübingen-Tagset (STTS[5]). Both models were deployed **locally** to adhere to the data privacy rules.

For Swiss Standard German transcriptions, we tested the statistical model de_core_news_sm provided by spaCy[6] as a supplement to the BERT-based models.

The models were chosen due to their availability before the data was collected.

Inter-annotator Agreement We recruited three native Swiss German speakers and three speakers with profound knowledge of Swiss Standard German with strong background in computational linguistics to annotate the gold-standard UPOS and STTS tags for sentences in the manual speech transcriptions. These annotations serve as ground truth for evaluating the performance of the three models discussed in Section 4.2. Prior to initiating the annotation process, we conducted training sessions with all annotators, during which the task instructions were explained in detail. We attached the instruction sheet in Appendix A for reference.

	Swiss German	Swiss Std. German
A&B	0.804	0.861
B&C	0.850	0.844
A&C	0.802	0.900

(a) IAA for UPOS tagging.

	Swiss German	Swiss Std. German
A&B	0.910	0.926
B&C	0.939	0.921
A&C	0.926	0.945

(b) IAA for STTS tagging.

Table 4: Pairwise linearly weighted Cohen's Kappa (Cohen, 1968) among three human annotators who are native Swiss German speakers for UPOS and STTS tagging on Swiss German (100 sentences) and Swiss Standard (Std.) German (80 sentences) transcriptions.

4.3. Morphological Features

German is a morphologically rich language that utilizes all 17 UPOS tags. For instances where a single POS tag is insufficient to capture lexical distinctions — such as the straightforward example of nouns in German, which can exhibit three different genders: masculine, feminine, and neutral—, features that describe these linguistic differences become essential.

Key	Value
Case	Acc, Nom, Gen, Dat
Number	Sing, Plur
Gender	Fem, Masc, Neut
Person	1, 2, 3
PronType	Art, Dem, Ind, Int, Prs, Rel
Mood	Ind, Sub, Imp
Tense	Past, Pres
VerbForm	Fin, Inf, Part
Definite	Def, Ind
Degree	Cmp, Pos, Sup
Foreign	Yes
Poss	Yes
Reflex	Yes

Table 5: Values for morphological categories. Notice that morphological categories rely on the spaCy model de_core_news_sm and are therefore language- and model-dependent.

To broaden the scope of our study, we annotated the morphology of language samples by assigning specific values to identified linguistic features (i.e., keys in our annotations). Table 5 provides a summary of the morphological values associated with these keys. Additionally, Table 6 outlines the mappings between POS tags and linguistic features, presented as key-value pairs.

The automatic prediction of morphological key-value features can be achieved using traditional statistical models (Can, 2011; Silfverberg and Lindén, 2011) or deep learning approaches (Tkachenko and Sirts, 2018; Bohnet et al., 2018; Klemen et al., 2023), both of which require a substantial amount of Swiss German language samples. However, deep learning approaches require a large amount of training data, which is currently absent in the Swiss LSA research.

To address this requirement, we are working on collecting additional language samples to support the training of these models. A comprehensive investigation of morphology prediction, however, is beyond the scope of this study and will be comprehensively addressed in future work.

5. Results

5.1. ASR Transcriptions

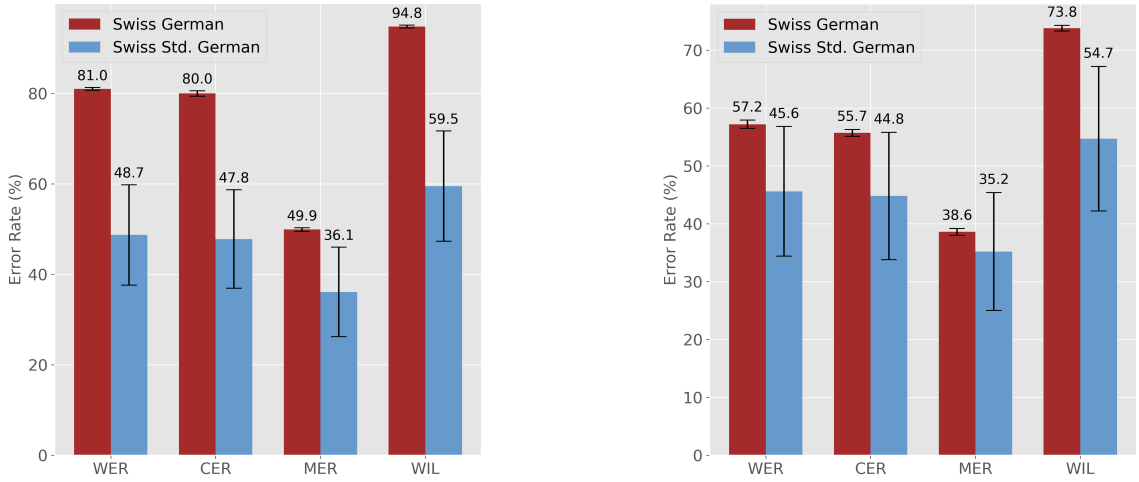
In Figure 2, we visualize the error rates of the Whisper-based ASR model transcribing the spontaneous speech recordings between one therapist and one child (Table 1). We report Word Error Rate (WER), Character Error Rate (CER), Match Error Rate (MER), and Word Information Lost (WIL) of Swiss German and Swiss Standard German tran-

[5]More details available at https://homepage.ruhr-uni-bochum.de/stephen.berman/Korpuslinguistik/Tagsets-STTS.html

[6]https://spacy.io/models/de, MIT licence

Tag	Key-Value Features
ADJA	{'Case': 'Acc', 'Degree': 'Pos', 'Gender': 'Fem', 'Number': 'Plur'}
ADJD	{'Degree': 'Pos'}
APPRART	{'Case': 'Dat', 'Gender': 'Neut', 'Number': 'Sing'}
ART	{'Case': 'Nom', 'Definite': 'Def', 'Gender': 'Neut', 'Number': 'Sing', 'PronType': 'Art'}
NN/NE	{'Case': 'Acc', 'Gender': 'Masc', 'Number': 'Sing'}
PDS/PDAT (plural)	{'Case': 'Nom', 'Number': 'Plur', 'PronType': 'Dem'}
PDS/PDAT (singular)	{'Case': 'Nom', 'Gender': 'Fem', 'Number': 'Sing', 'PronType': 'Dem'}
PIS	{'Gender': 'Neut', 'PronType': 'Ind'}
PIAT	{'Case': 'Nom', 'Gender': 'Fem', 'Number': 'Sing', 'PronType': 'Ind'}
PPER (other)	{'Case': 'Nom', 'Number': 'Sing', 'Person': '3', 'PronType': 'Prs'}
PPER (singular)	{'Case': 'Nom', 'Gender': 'Neut', 'Number': 'Sing', 'Person': '3', 'PronType': 'Prs'}
PPOSAT	{'Case': 'Acc', 'Gender': 'Fem', 'Number': 'Sing', 'Poss': 'Yes', 'PronType': 'Prs'}
PPOSS	{'Case': 'Nom', 'Gender': 'Masc', 'Number': 'Sing', 'Poss': 'Yes', 'PronType': 'Prs'}
PRF	{'Case': 'Acc', 'Number': 'Sing', 'Person': '3', 'PronType': 'Prs', 'Reflex': 'Yes'}
PRELS/PRELAT	{'Case': 'Acc', 'Gender': 'Neut', 'Number': 'Sing', 'PronType': 'Rel'}
PWS	{'Case': 'Nom', 'Gender': 'Masc', 'Number': 'Sing', 'PronType': 'Int'}
PWAV	{'PronType': 'Int'}
VERB	{'Mood': 'Ind', 'Number': 'Sing', 'Person': '3', 'Tense': 'Pres', 'VerbForm': 'Fin'}
VERB IMP	{'Number': 'Sing'}
VERB INF	{'VerbForm': 'Inf'}
VVPP	{'VerbForm': 'Part'}

Table 6: Morphological features for STTS tags (based on `spaCy`), sorted alphabetically. The UPOS tags are converted to STTS tags using a conversion look-up table.



(a) Original Transcriptions.

(b) Normalized Transcriptions.

Figure 2: Average ASR results on Whisper transcriptions with standard deviations.

scriptions.[7] The WER for Swiss German is 81%, and 48.7% for Standard German. After normalizing the transcripts, the WER is 57.2% for Swiss German and for Standard German 45.6%

5.2. Inter-annotator Agreement on Human Labelling of POS Tags

Inter-annotator agreement (IAA) scores were calculated based on 100 sentences for Swiss German and 80 sentences for Swiss Standard German, as one annotator for Swiss Standard German was unable to complete annotations for all 100 sentences. These sentences were randomly sampled from the corresponding transcriptions.

Overall, we achieved high IAA scores for both UPOS tagging (above 0.8) and STTS tagging (above 0.9), as measured using linearly weighted Cohen's Kappa (Cohen, 1968). The detailed scores are presented in Table 4.

5.3. Automatic POS Tagging

In Figure 3, we present the POS tagging performance of the BERT-based models and the `spaCy`

[7] We use the `JiWER` Python package for all error rate calculations, Apache-2.0 license.

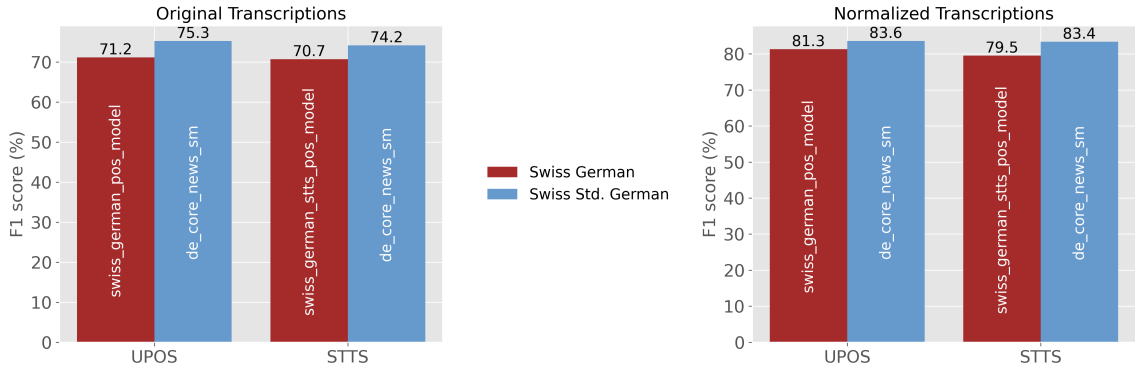


Figure 3: POS tagging results with BERT-based POS tagging model and `spaCy` model, measured on all transcription data for Swiss German and Swiss Standard German.

model, evaluated on the complete transcription datasets for both Swiss German and Swiss Standard German. All models achieved F1 scores above 70, with significantly higher results observed on normalized transcriptions (above or nearly 80) compared to original transcriptions.

6. Discussion

6.1. Challenges of Manual and Automatic LSA

LSA poses many challenges despite being a gold standard tool in speech therapy (Klatte et al., 2022; Bawayan et al., 2022), which have to be taken into account in the manual use as well as in the development of a semi-automatic pipeline for LSA. Factoring them in from the first steps of development as well as searching for solutions is part of ongoing and future research.

Data collection. Speech and language samples of children are typically recorded in naturalistic environments, such as in free conversations, which include both relevant speech and background noise like TV soundtracks (Lüdtke et al., 2023), or, in our case, noise of other children playing and street noise in the background. Extracting meaningful speech utterances and transcribing them into text units which speech-language pathologists can directly work with, is often a challenging task.

Manual annotation. Linguistic features such as errors on the levels of syntax, morphology, semantics, or phonology are often manually annotated and analysed by speech-language pathologists, which due to its time-intensive nature makes the use in therapeutic assessment less common (Owens Jr et al., 2018).

Generalization difficulty. Language error patterns (such as wrong word order) vary from one child to another and from dialect to dialect, which limits the generalizability of LSA methods.

Use of LLMs In recent years, large language models (LLMs) have undergone significant advancements. However, querying commercial LLMs such as ChatGPT with research data, including language samples from children, is not compliant with Swiss data protection regulations. Furthermore, the use of commercial LLMs for language sample analysis (LSA) is neither ethical nor practical due to the highly sensitive nature of children's speech transcriptions and the risk of data contamination. As a result, progress in automating LSA for the diagnosis of developmental language disorder (DLD) remains limited and understudied.

Swiss German The linguistic landscape of Switzerland presents unique challenges for applying NLP methods (Parida et al., 2020). The prevalence of Swiss German dialects, which differ significantly from Standard German in their linguistic structure, complicates real-world NLP practice. Automatic speech recognition (ASR) on Swiss German typically produces Standard German transcriptions, thereby omitting or distorting important dialectal information or not being able to represent Swiss German faithfully, for example sentences such as "Ich gang go poste." (*I'm going shopping.*) can be translated into "Ich gehe einkaufen.", losing the "go", which does not exist in German but is essential in Swiss German. Models trained on written Standard German typically perform poorly on Swiss German due to the lack of a standardized written form (Kew et al., 2020; Nigmatulina et al., 2020). Despite this, written Swiss German is increasingly used in digital communication, such as social media and online messaging, where individual writing styles introduce further variation

(Hollenstein and Aepli, 2014). In educational settings, children use Swiss Standard German. Swiss Standard German is a variety of Standard German but still contains words ("helvetisms") and grammar rules that render it different from Standard German of Germany or Austria. Many children who do not speak Swiss German as their native language primarily, or exclusively, communicate in Swiss Standard German. This highlights the necessity of NLP methods that can effectively process both Swiss Standard German and Swiss German while accommodating the inherent variation within Swiss German dialects.

Research in other languages such as English supports the effectiveness of LSA in diagnosing DLD (Ramos et al., 2022) and the effectiveness of using automated LSA (Miller et al., 1985; Pye, 1994). However, relevant studies are less present for German data and entirely missing for Swiss German. Our preliminary study investigates the potential of using NLP methods as a step closer to working with Swiss German data.

6.2. ASR Transcriptions

The majority of transcription errors committed by the Whisper model can be attributed to child speech. Specifically, Whisper often failed to recognize the children's utterances (but not the utterances of the adult therapists) or generated repeated words. After orthographically correcting the manual transcriptions (i.e., normalization; for examples, see columns **word** (original) and **normalized** in Appendix E), the error rates were significantly reduced (see Figure 2b). This highlights the continued need for normalization of speech transcriptions in practice to address the limitations of ASR models. This problem could also potentially be mitigated by fine-tuning the models with specific methods of transcription, which would allow keeping the important transcribed information as well as reaching a sufficient quality of transcription.

In our case study, indicated by the less prominent results of the Whisper transcriptions, fully relying on ASR transcriptions was not meaningful at this stage due to the limited availability of children's speech data in Swiss German, which precluded fine-tuning even the small variant of the Whisper model. This highlights the importance of both naturally expanding the dataset and exploring alternative approaches, such as data augmentation via speech synthesis using generative models (Ren et al., 2021; Ao et al., 2022; Toyin et al., 2024), to address the challenges of processing underrepresented and atypical languages in a speech-language pathology context.

6.3. Inter-annotator Agreement on Human Labelling of POS Tags

Our results indicate a strong level of agreement, demonstrating the reliability of our annotation process. Prior research on IAA for German and Latin POS tagging (Brants, 2000; Stüssi and Ströbel, 2024) has reported even higher IAA scores with professionally trained annotators, which suggests that the annotation quality could be further improved with expert training.

We argue that three primary factors contribute to the differences observed in our study: (1) While no major discrepancies exist, Swiss Standard German differs from the Standard German used in Germany, which can lead to disagreements among annotators; (2) Swiss German lacks standardized spelling, and our annotators originate from different Swiss German-speaking cantons, resulting in variations in their interpretation of syntactic functions; and (3) Due to the presence of erroneous, incomplete, or atypical sentence structures produced by children with DLD, some ambiguity remained, leading to variability in interpretation among annotators. In Appendix B, we provide examples of common discrepancies between our annotators.

6.4. Automatic POS Tagging

The results for the more general statistical `spaCy` model on Swiss Standard German data are consistently outperforming the two BERT-based POS tagging models specifically trained on UPOS and STTS for Swiss German annotations on the Swiss German data, further highlighting the difficulties specific to Swiss German in NLP methods. All in all, these findings highlight that the models used in this study can perform reasonably well on Swiss German and Swiss Standard German transcriptions while offering the advantages of being significantly less computationally expensive and more ethical in their deployment compared to commercial LLMs. More Swiss German training data in general, as well as finetuning the model to the specifics of our dataset (e.g. children's' speech and atypical speech) is expected to improve the performance.

6.5. Current Challenges

Despite achieving a reasonably good performance in automatic speech transcription and POS tagging, it is important to emphasize that the current results remain insufficient for therapeutic practice. This limitation is primarily due to the persistent need for manual correction of speech transcriptions. Since diagnostic applications demand highly precise speech transcription and linguistic analysis, there is strong motivation to further enhance NLP

approaches not based on commercial LLMs, such as by fine-tuning existing BERT-based POS tagging models with additional data and developing more advanced ASR models for Swiss German speech.

In Appendix D, we present our prototype software developed specifically for speech-language pathologists. As we continue to gather user feedback from real-world practice, we adhere to human-in-the-loop design principles and plan to enhance the software with more rigorously validated features for better user experience.

7. Conclusion and Future Work

In this study, we have addressed the challenges of automating key steps in language sample analysis by employing non-commercial NLP models for ASR and POS tagging. We evaluated the zero-shot capabilities of these models on both tasks and reported the empirical performance. Our work underscores the feasibility of employing ethical NLP approaches not based on commercial LLMs in the setting of speech-language pathology, particularly when handling sensitive data such as children's speech. In future work, we aim to expand the dataset by incorporating more diverse samples of children's speech with and without DLD. Additionally, we plan to develop specialized ASR and POS tagging models tailored to Swiss German of children and evaluate whether analyses based on automatically created transcripts and automated POS tagging can reliably predict DLD in Swiss German-speaking children. We also plan to expand the linguistic analyses in broader contexts, especially on the syntactic level (such as dependency parsing, see the column **dependency** in Table 20 in Appendix E), ultimately facilitating the semi-automatic diagnosis of DLD for children in Switzerland.

8. Limitations

The primary limitations of our work are as follows: (1) Due to the data sparsity and difficulty of data acquisition, our sample size is relatively low compared to evaluations in contexts other than the speech and language disorder context; (2) We did not conduct further investigations into morphological features, as, to the best of our knowledge, no existing NLP approaches not based on commercial LLMs for morphological prediction in Swiss German are currently available; (3) We did not benchmark the evaluation with locally deployed LLMs of the newest generation as this will be part of a future study; (4) All Swiss German ASR models known to us transcribe the text directly into Standard German. However, for our application, it would be beneficial for the text to be in Swiss German. There is not yet any research showing that DLD approaches for other languages can be effectively applied to diagnosing DLD of Swiss German-speaking children. For this reason, the usefulness of the Standard German transcriptions obtained is limited; (5) Automatic tools for diagnosing DLD in children are not infallible; they should only be used in combination with other screening methods and by speech-language pathologists.

9. Ethical Statement

Our study received ethical approval from the responsible university's research committee. Informed consent was obtained from the parents of all participating children and the speech-language pathologists, allowing for the recording, processing, storage, and controlled sharing of data. To ensure accessibility, the consent form was provided in simplified language to facilitate understanding for parents. Children were informed about the study and provided their oral consent to participate. None of the collected data have been processed through any commercial LLMs. All data processing was conducted locally. Permission was granted to publish the data from 41 recordings in a public repository. In addition to the small subset used for this paper, the whole dataset will be released in the near future.

10. Acknowledgement

This work was supported by the Foundation Special Education Centre Fribourg (Stiftung Heilpädagogisches Zentrum Fribourg) and the Foundation for Speech Therapy in the Canton of Zurich (Förderstiftung für das Sprachheilwesen im Kanton Zürich). We thank Susanne Kempe-Preti, Pascale Schaller, Julia Winkes, Sonja Schäli, Lena Graf, Ramona Rüegg, Lukas Fischer, Anne Göhring, Michelle Wastl and Angela Heldstab for their valuable input to the study.

11. Bibliographical References

Noëmi Aepli and Rico Sennrich. 2022. Improving Zero-shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-level Noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing

Li, Yu Zhang, et al. 2022. SpeechT5: Unified-modal Encoder-decoder Pre-training for Spoken Language Processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

Yuriy Arabskyy, Aashish Agarwal, Subhadeep Dey, and Oscar Koller. 2021. Dialectal Speech Recognition and Translation of Swiss German Speech to Standard German Text: Microsoft's Submission to SwissText 2021. *arXiv preprint arXiv:2106.08126*.

Rebecca Bawayan, Jennifer A Brown, Rebecca Bawayan, and Jennifer A Brown. 2022. Language Sample Analysis Consideration and Use: A Survey of School-based Speech Language Pathologists. *Clinical Archives of Communication Disorders*, 7(1):15–28.

Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. 2022. Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *Applied Sciences*, 12(9):4419.

Dorothy VM Bishop. 2006. What Causes Specific Language Impairment in Children? *Current directions in psychological science*, 15(5):217–221.

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652.

Thorsten Brants. 2000. Inter-annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*.

Burcu Can. 2011. *Statistical Models for Unsupervised Learning of Morphology and POS Tagging*. Ph.D. thesis, University of York.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological bulletin*, 70(4):213.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift*. Cornelsen Schweiz AG.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel Multidialectal Corpus of Spoken Swiss German. *arXiv preprint arXiv:2103.11401*.

Eyal Liron Dolev, Clemens Fidel Lutz, and Noëmi Aepli. 2024. Does Whisper Understand Swiss German? An Automatic, Qualitative, and Human Evaluation. *arXiv preprint arXiv:2404.19310*.

Keyur Gabani, Thamar Solorio, Yang Liu, Khairunnisa Hassanali, and Christine A Dollaghan. 2011. Exploring A Corpus-based Approach for Detecting Language Impairment in Monolingual English-speaking Children. *Artificial Intelligence in Medicine*, 53(3):161–170.

John F Gallagher and Jill R Hoover. 2020. Measure What You Treat: Using Language Sample Analysis for Grammatical Outcome Measures in Children with Developmental Language Disorder. *Perspectives of the ASHA Special Interest Groups*, 5(2):350–363.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2012. Evaluating NLP Features for Automatic Prediction of Language Impairment Using Child Speech Transcripts. In *INTERSPEECH*, pages 1339–1342.

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.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of Whisper Models to Child Speech Recognition. *arXiv preprint arXiv:2307.13008*.

Tannon Kew, Iuliia Nigmatulina, Lorenz Nagele, and Tanja Samardzic. 2020. UZH TILT: A Kaldi Recipe for Swiss German Speech to Standard German Text. In *SwissText/KONVENS*.

Inge S Klatte, Vera Van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language Sample Analysis in Clinical Practice: Speech-language Pathologists' Barriers, Facilitators, and Needs. *Language, Speech, and Hearing Services in Schools*, 53(1):1–16.

Matej Klemen, Luka Krsnik, and Marko Robnik-Šikonja. 2023. Enhancing Deep Neural Networks with Morphological Information. *Natural Language Engineering*, 29(2):360–385.

Dancheng Liu, Jason Yang, Ishan Albrecht-Buehler, Helen Qin, Sophie Li, Yuting Hu, Amir Nassereldine, and Jinjun Xiong. 2024. Automatic

Screening for Children with Speech Disorder using Automatic Speech Recognition: Opportunities and Challenges. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 308–313.

Ulrike Lüdtke, Juan Bornman, Febe De Wet, Ulrich Heid, Jörn Ostermann, Lars Rumberg, Jeannie Van der Linde, and Hanna Ehlert. 2023. Multidisciplinary Perspectives on Automatic Analysis of Children's Language Samples: Where Do We Go From Here? *Folia Phoniatrica et Logopaedica*, 75(1):1–12.

Carina Lüke, Christina Kauschke, Andrea Dohmen, Andrea Haid, Christina Leitinger, Claudia Männel, Tanja Penz, Steffi Sachse, Wiebke Scharff Rethfeldt, Julia Spranger, et al. 2023. Definition and Terminology of Developmental Language Disorders—interdisciplinary Consensus Across German-speaking Countries. *Plos one*, 18(11):e0293736.

Jon Miller, Robin Chapman, et al. 1985. Systematic Analysis of Language Transcripts. *Madison, WI: Language Analysis Laboratory*.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for Non-standardised Languages with Dialectal Variation: the Case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24.

Robert E Owens Jr, Stacey L Pavelko, and Dennis Bambinelli. 2018. Moving Beyond Mean Length of Utterance: Analyzing Language Samples to Identify Intervention Targets. *Perspectives of the ASHA Special Interest Groups*, 3(1):5–22.

Shantipriya Parida, Esaú Villatoro-Tello, Sajit Kumar, Petr Motlicek, and Qingran Zhan. 2020. Idiap Submission to Swiss-German Language Detection Shared Task. In *SwissText/KONVENS*, page 7.

Michel Plüss, Jan Milan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Huerlimann, 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.

Michel Plüss, Manuela Huerlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, et al. 2022. DS-200: A Swiss German Speech to Standard German Text Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2020. Swiss Parliaments Corpus, An Automatically Aligned Swiss German Speech to Standard German Text Corpus. *arXiv preprint arXiv:2010.02810*.

Niclas Pokel, Pehuén Moure, Roman Boehringer, and Yingqiang Gao. 2025. Adapting Foundation Speech Recognition Models to Impaired Speech: A Semantic Re-chaining Approach for Personalization of German Speech. In *12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)*, pages 82–86.

Alexandros Potamianos, Shrikanth S Narayanan, and Sungbok Lee. 1997. Automatic Speech Recognition for Children. In *Eurospeech*, volume 97, pages 2371–2374.

Clifton Pye. 1994. The CHILDES Project: Tools for Analyzing Talk.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-scale Weak Supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Michelle N Ramos, Penelope Collins, and Elizabeth D Peña. 2022. Sharpening Our Tools: A Systematic Review to Identify Diagnostically Accurate Language Sample Measures. *Journal of Speech, Language, and Hearing Research*, 65(10):3890–3907.

Michelle Nichols Ramos. 2024. *Using Language Sample Analysis to Identify Developmental Language Disorder in Bilingual Children*. Ph.D. thesis, University of California, Irvine.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and High-quality End-to-end Text to Speech. In *International Conference on Learning Representations*.

Lars Rumberg, Hanna Ehlert, Ulrike Lüdtke, and Jörn Ostermann. 2021. Age-Invariant Training for End-to-end Child Speech Recognition Using Adversarial Multi-task Learning. In *Interspeech*, pages 3850–3854.

Alessandra Sansavini, Maria Elena Favilla, Maria Teresa Guasti, Andrea Marini, Stefania Millepiedi, Maria Valeria Di Martino, Simona Vecchi, Nadia Battajon, Laura Bertolo,

Olga Capirci, Barbara Carretti, Maria Paola Colatei, Cristina Frioni, Luigi Marotta, Sara Massa, Letizia Michelazzo, Chiara Pecini, Silvia Piazzalunga, Manuela Pieretti, Pasquale Rinaldi, Renata Salvadorini, Cristiano Termine, Mariagrazia Zuccarini, Simonetta D'Amico, Anna Giulia De Cagno, Maria Chiara Levorato, Tiziana Rossetto, and Maria Luisa Lorusso. 2021. Developmental Language Disorder: Early Predictors, Age for the Diagnosis, and Diagnostic Tools. A Scoping Review. *Brain Sciences*, 11(5).

Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. 2022. Swiss German Speech to Text System Evaluation. *arXiv preprint arXiv:2207.00412*.

Miikka Silfverberg and Krister Lindén. 2011. Combining Statistical Models for POS Tagging using Finite-state Calculus. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 183–190.

Daniel V Smith, Alex Sneddon, Lauren Ward, Andreas Duenser, Jill Freyne, David Silvera-Tawil, and Angela Morgan. 2017. Improving Child Speech Disorder Assessment by Incorporating Out-of-domain Adult Speech. In *Interspeech*, pages 2690–2694.

Thamar Solorio. 2013. Survey on Emerging Research on the Use of Natural Language Processing in Clinical Language Assessment of Children. *Language and Linguistics Compass*, 7(12):633–646.

Samuel Stucki, Mark Cieliebak, and Jan Deriu. 2025. SwissGPC v1. 0–The Swiss German Podcasts Corpus. *arXiv preprint arXiv:2509.19866*.

Elina Stüssi and Phillip Ströbel. 2024. Part-of-speech Tagging of 16th-Century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206.

Vincenzo Timmel, Claudio Paonessa, Reza Kakooee, Manfred Vogel, and Daniel Perruchoud. 2024. Fine-tuning Whisper on Low-Resource Languages for Real-World Applications. *arXiv preprint arXiv:2412.15726*.

Alexander Tkachenko and Kairit Sirts. 2018. Modeling Composite Labels for Neural Morphological Tagging. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 368–379.

J Bruce Tomblin, Nancy L Records, and Xuyang Zhang. 1996. A System for the Diagnosis of Specific Language Impairment in Kindergarten Children. *Journal of Speech, Language, and Hearing Research*, 39(6):1284–1294.

Hawau Toyin, Hao Li, and Hanan Aldarmaki. 2024. STTATTS: Unified Speech-to-text And Text-to-speech Model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6853–6863.

Vivian van Wijngaarden, Hester de Wilde, Dieuwke Mink van der Molen, Jildo Petter, Inge Stegeman, Ellen Gerrits, Adriana L Smit, and Marie-José van den Boogaard. 2024. Genetic Outcomes in Children with Developmental Language Disorder: A Systematic Review. *Frontiers in pediatrics*, 12:1315229.

A. Instructions for Human Annotation of Part-of-speech Tagging on Gold Transcriptions (translated from the original German file)

Task Description. Your task is to annotate a small dataset of spontaneous speech sentences of children with typical and atypical language development in Swiss German and Swiss Standard German. You received an Excel sheet with around 100 sentences from different children (K as *Kind* (child) in German) and professionals (FP as *Fachperson* (specialist) in German) interviewing the children in either Swiss German or Swiss Standard German. These sentences are randomly selected from different gold transcription files. In the first column you find the human transcriptions, while all other columns are generated automatically and need to be corrected. Please correct the annotations of:

- Part-of-speech tagging with the UPOS tags (more information here[1]).
- Part-of-speech tagging with the STTS tags (more information here[2]).
- Morphology for certain part-of-speech tags.
- Subject-verb agreement (SVA).

Annotation Process. Please adhere to the following steps:

1. Read this manual carefully.
2. Read the documentation of the tag sets.
3. Perform trial annotations using the control sentences constructed individually and send them back.
4. If you pass the trial annotations, you will receive real annotation sentences.
5. During the annotation, please keep in mind that:
    - For the annotation you can use whatever assistance you want (such as look up tables, Duden, the internet, etc.);
    - Please **do not use the `spaCy` model `de_core_news_sm` as well as the `swiss_german_stts_pos_model` and the `swiss_german_pos_model` models from Huggingface**, as they are baselines of our study;
    - Please **do not copy and paste the sentences into commercial LLMs**, as we must respect the data protection policy of Switzerland;
    - The sentences are random samples of different transcriptions. Please use only the current sentence as context;
    - You can annotate in whatever order you like.
6. For morphology, annotate as much as you can as long as it is determinable. If something is not determinable, please leave it out blank.
7. For subject-verb agreement:
    - Please mark all conjugated verbs with 'v';
    - Please mark the main part that determines the verb form (subjects) as 'sb';
    - For the contracted forms of Swiss German (such as "gehen wir" → "gömmer", "kann er sie entsorgen" → "chanerse entsorge"), use sb_v or v_sb, depending on the order in the contraction.
8. Leave `<sentence>`, the sentence separator, as empty.
9. Tag all `UNK` as X/XY.
10. Tag all `NAME` (anonymized name) as PROPN/NE.
11. Use the tag PROAV instead of PAV (adhering to `spaCy`).

[1] https://universaldependencies.org/u/pos/

[2] https://homepage.ruhr-uni-bochum.de/stephen.berman/Korpuslinguistik/Tagsets-STTS.html

12. In STTS, the verb `sein` is always tagged as VAxxx, even when used as a full verb. However, UPOS distinguishes between the functions of the verb: when a verb is used as a full verb, can can thus be replaced by `sich befinden`, it is tagged as VERB.

13. The STTS tag set labels all occurrences of auxiliary verbs (`sein, werden, haben`) as well as copula with VAxxx, modal verbs (`müssen, dürfen, wollen, mögen, können, sollen`) are labelled with VMxxx.

14. Ja/Nein: Please annotate as PTKANT if

    - Standing alone;
    - Is part of an answer;
    - Used as a query (Rückfrage).

    Except when used as modal particle (in German "Abtönungspartikel", for instance, "Das ist ja wirklich schön"), annotate as ADV.

15. Interjections encompass:

    - All signals of understanding (in German "Verständigungssignale", for instances, "oh', "ah", "aha", "wow", "hmm");
    - All onomatopoeias (for instances, "brummbrumm", "gluglug", "miau", "wuff").

16. Corrections, word fragments, errors (which are then corrected, for instance, "eine Bri-Brille"), please annotate as X/XY.

17. If a child made a grammatical error and used the wrong form, **please annotate what was exactly said and not what it should be**, as we want to identify these errors later. If not all information can be clearly determined from the used word itself, assume the correct form.

18. For cases that are hard to tag (e.g., the child used irregular forms, wrong or fantasy words, etc.), please do your best.

Rules of Swiss German. Most rules are identical between Swiss German and Standard German. However, there are still some linguistic differences. Please read the chapter *Differences to German UD Guidelines* here[3] as a reference.

Conversion between UPOS tag set and STTS tag set. See more information here[4]. Please be aware that the conversion does not apply to all cases.

[3] https://universaldependencies.org/gsw/
[4] https://universaldependencies.org/tagset-conversion/de-stts-uposf.html

B. Canonical Examples of POS Tagging Disagreements between Human Annotators

B.1. Examples in Swiss German

Disagreement 1 Whether a word is a proper noun (PROPN), noun (NOUN), or foreign word (X).

	ebe	genau	ja	de	het	der	paw	patrol	gfalle	?
A	ADV	ADV	PART	ADV	AUX	PRON	*X*	*X*	VERB	PUNCT
B	ADV	ADV	PART	ADV	AUX	PRON	*PROPN*	*PROPN*	VERB	PUNCT
C	ADV	ADV	PART	ADV	AUX	PRON	*PROPN*	*PROPN*	VERB	PUNCT

Table 7: Example sentence corresponding to the English translation *so exactly then you liked paw patrol?*

Disagreement 2 Whether a word is an adverb (ADV) or, e.g., an adjective (ADJD) or interjection (ITJ).

	ja	hm	genau	ah	schpannend
A	PTKANT	ITJ	*ADJD*	ITJ	ADJD
B	PTKANT	ITJ	*ADV*	ITJ	ADJD
C	PTKANT	ITJ	*ADV*	ITJ	ADJD

Table 8: Example sentence corresponding to the English translation *yes hm exactly ah interesting*.

	gäll	was	dänksch	was	isch	das
A	*ITJ*	PWS	VVFIN+	PWS	VAFIN	PDS
B	*ADV*	PWS	VVFIN	PWS	VAFIN	PDS
C	*ITJ*	PWS	VVFIN+	PWS	VAFIN	PDS

Table 9: Example sentence corresponding to the English translation *what do you think what this is?*

Disagreement 3 Distinguishing different types of pronouns (e.g., PDS, PDAT).

	gend	recht	gas	uf	dene	trotti
A	VVFIN	ADV	NN	APPR	*PDS*	NN
B	VVFIN	ADV	PTKVZ	APPR	*PDAT*	NN
C	VVFIN	ADV	NN	APPR	*PDS*	NN

Table 10: Example sentence corresponding to the English translation *they really step on the gas on these scooters*

Disagreement 4 Whether to tag attributive pronouns (PPOSAT in STTS) as determiner (DET) or pronoun (PRON) in UPOS.

	oh	verzell	mal	wy	fyrsch	dù[5]	din	gebùrtstag	?
A	INTJ	VERB	ADV	ADV	VERB	PRON	*DET*	NOUN	PUNCT
B	INTJ	VERB	ADV	CCONJ	VERB	PRON	*PRON*	NOUN	PUNCT
C	INTJ	VERB	ADV	ADV	VERB	PRON	*DET*	NOUN	PUNCT
STTS	ITJ	VVIMP	ADV	PWAV	VVFIN	PPER	*PPOSAT*	NN	$.

Table 11: Example sentence corresponding to the English translation *oh tell me how do you celebrate your birthday?*

[5]Adding an accent symbol to a vowel describes its quality adapted from Dieth (1986).

Disagreement 5 Whether a word is a concatenation of multiple words (e.g., VVFIN+) or not (e.g., VVFIN) (especially in the case of a verb in second person singular, as it can stand without PPER in Swiss German).

	gäll	was	dänksch	was	isch	das
A	ITJ	PWS	*VVFIN+*	PWS	VAFIN	PDS
B	ADV	PWS	*VVFIN*	PWS	VAFIN	PDS
C	ITJ	PWS	*VVFIN+*	PWS	VAFIN	PDS

Table 12: Example sentence corresponding to the English translation *what do you think this is?*

B.2. Examples in Swiss Standard German

Disagreement 1 Whether a word is nominalized (used as a noun) or not.

	aber	du	hast	recht	das	macht	man	doch	eigentlich	mit	einem	stock
A	CCONJ	PRON	AUX	*NOUN*	PRON	VERB	PRON	ADV	ADV	ADP	DET	NOUN
B	CCONJ	PRON	AUX	*NOUN*	PRON	VERB	PRON	ADV	ADV	ADP	DET	NOUN
C	CCONJ	PRON	AUX	*ADV*	PRON	VERB	PRON	ADV	ADV	ADP	DET	NOUN

Table 13: Example sentence corresponding to the English translation *but you are right actually you do this with a stick*.

Disagreement 2 For incomplete or erroneous words, what is the right way to interpret it (the correct word in the first example should be “meinte” (“meant”) as one word, and in the second example, “warte” (“wait”).

	mein	te	an
A	*PPOSAT*	*NN*	PTKVZ
B	*PPOSAT*	*XY*	XY
C	*VVFIN*	*VVFIN*	APPR

(a) English translation of the example sentence: *mean-ed on*.

	wate	an	freitag
A	*NN*	APPR	NN
B	*XY*	APPR	NN
C	*VVIMP*	APPR	NN

(b) Example sentence corresponding to the English translation *wait on friday*.

Table 14: Comparative morphological annotation examples from two German utterances.

Disagreement 3 Whether a verb is used as auxiliary (VA) or full verb (VV) (differences exist between the two tag sets).

	ja	diese	zeitung	ich	hab	das	gerne	aber	ich	hat	ein	zeichnung
A (UPOS)	PART	DET	NOUN	PRON	VERB	PRON	ADV	CCONJ	PRON	*AUX*	DET	NOUN
A (STTS)	PTKANT	PDAT	NN	PPER	VAFIN	PDS	ADV	KON	PPER	*VAFIN*	ART	NN
B (UPOS)	PART	DET	NOUN	PRON	VERB	PRON	ADV	CCONJ	PRON	*AUX*	DET	NOUN
B (STTS)	PTKANT	PDAT	NN	PPER	VAFIN	PDS	ADJD	KON	PPER	*VAFIN*	ART	NN
C (UPOS)	PART	DET	NOUN	PRON	VERB	PRON	ADV	CCONJ	PRON	*VERB*	DET	NOUN
C (STTS)	PTKANT	PDAT	NN	PPER	VAFIN	PDS	ADV	KON	PPER	*VAFIN*	ART	NN

Table 15: Example sentence corresponding to the English translation *yes this newspaper I like this but I has a drawing*.

C. Per-speaker-group ASR and POS Tagging Results

Observation 1 Whisper has significantly less ASR errors on transcriptions of speech-language pathologists than of children.

Original Transcripts	WER ↓		CER ↓		MER ↓		WIL ↓	
	FP	K	FP	K	FP	K	FP	K
Swiss German	0.772	0.860	0.469	0.556	0.761	0.850	0.927	0.970
Swiss Std. German	0.377	0.600	0.278	0.445	0.370	0.591	0.463	0.722

Table 16: ASR results evaluated on **original** speech transcriptions of speech-language pathologists (FP) and children (K).

Normalized Transcripts	WER ↓		CER ↓		MER ↓		WIL ↓	
	FP	K	FP	K	FP	K	FP	K
Swiss German	0.506	0.681	0.337	0.466	0.494	0.664	0.666	0.840
Swiss Std. German	0.365	0.550	0.275	0.430	0.359	0.541	0.444	0.651

Table 17: ASR results evaluated on **normalized** speech transcriptions of speech-language pathologists (FP) and children (K).

Observation 2 POS tagging models have higher F1 scores on transcriptions of speech-language pathologists (i.e. adults) than of children.

Original Transcriptions	UPOS (F1 Score ↑)		STTS (F1 Score ↑)	
	FP	K	FP	K
Swiss German	0.735	0.673	0.728	0.675
Swiss Std. German	0.844	0.659	0.831	0.654

Table 18: POS tagging results on **original** speech transcriptions of speech-language pathologists (FP) and children (K).

Observation 3 Swiss Standard German transcriptions achieved generally higher evaluation results than Swiss German transcriptions.

Normalized Transcriptions	UPOS (F1 Score ↑)		STTS (F1 Score ↑)	
	FP	K	FP	K
Swiss German	0.835	0.787	0.823	0.762
Swiss Std. German	0.878	0.796	0.852	0.807

Table 19: POS tagging results on **normalized** speech transcriptions of speech-language pathologists (FP) and children (K).

Observation 4 Orthographic normalization of speech transcriptions helps in boosting the performance of both ASR model and POS tagging model.

D. The LSA Software for Therapeutic Practice of DLD Diagnosis

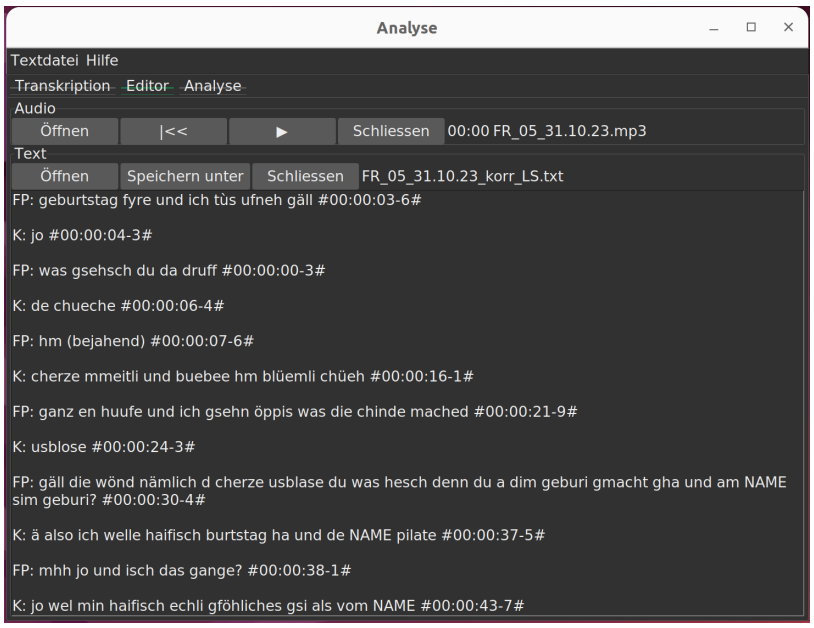


Figure 4: Editor of the software. After automatically transcribing the recordings, the transcript is opened in the editor. Here, the recordings can be played and the transcript can be corrected.

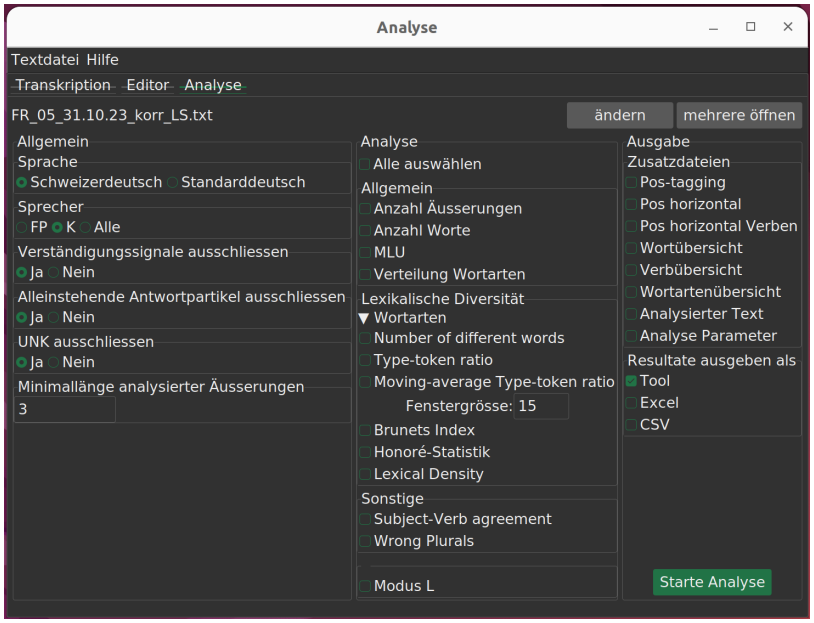


Figure 5: Analysis: After correcting the automatic transcription manually, the analysis can be started. Providing different options for the analysis, such as which speakers should be analyzed and what should be filtered out, a personalized analysis can be executed, containing values such as mean length of utterance, distribution of POS tags, and subject verb agreement as well as additional files with overviews of all verbs and POS tagging.

E. Annotation Examples

send_id	speaker	word_id	word	normalized	lemma	UPOS tag	STTS tag	morphology	SVA	dependency
62	FP	0	warst	Warst	sein	VERB	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘2’, ‘Tense’: ‘Past’, ‘VerForm’: ‘Fin’}	v	ROOT
62	FP	1	du	du	du	PRON	PPER	{‘Case’: ‘Norm’, ‘Number’: ‘Sing’, ‘Person’: ‘2’, ‘PronType’: ‘Prs’}	sb	sb
62	FP	2	diesen	diesen	dieser	DET	PDAT	{‘Case’: ‘Acc’, ‘Gender’: ‘Masc’, ‘Number’: ‘Sing’, ‘PronType’: ‘Dem’}		nk
62	FP	3	sommer	Sommer	Sommer	NOUN	NN	{‘Case’: ‘Acc’, ‘Gender’: ‘Masc’, ‘Number’: ‘Sing’}		oa
62	FP	4	auch	auch	auch	ADV	ADV			mo
62	FP	5	schon	schon	schon	ADV	ADV			mo
62	FP	6	in	in	in	ADP	APPR			mo
62	FP	7	der	der	der	DET	ART	{‘Case’: ‘Dat’, ‘Definite’: ‘Def’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’, ‘PronType’: ‘Art’}		nk
62	FP	8	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Dat’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}		oa
62	FP	9	?	?	?	PUNCT	$.			punct
63	K	0	ähm	Ähm	ähm	INTJ	ITJ			mo
63	K	1	ja	ja	ja	PART	PTKANT			mo
63	K	2	abeR	aber	aber	CCONJ	KON			mo
63	K	3	is	ist	sein	AUX	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	sb
63	K	4	de	der	der	PRON	PDS	{‘Case’: ‘Nom’, ‘Gender’: ‘Masc’, ‘Number’: ‘Sing’, ‘PronType’: ‘Dem’}	sb	uc
63	K	5	mine	meine	mein	PRON	PPOSAT	{‘Case’: ‘Nom’, ‘Gender’: ‘Neut’, ‘Number’: ‘Sing’, ‘Poss’: ‘Yes’, ‘PronType’: ‘Prs’}		nk
63	K	6	haus	Haus	Haus	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Neut’, ‘Number’: ‘Sing’}		ams
63	K	7	zün	zün	zün	X	XY			pd
63	K	8	ist	ist	sein	VERB	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	ROOT
63	K	9	da	da	da	ADV	ADV			mo
63	K	10	drauf	drauf	drauf	ADV	PROAV			mo
63	K	11	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}	sb	oc
63	K	12	kann	kann	können	AUX	VMFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	sb
63	K	13	kann	kann	können	AUX	VMFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	sb
63	K	14	türe	Türe	Tür	NOUN	NN	{‘Case’: ‘Acc’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}		sb
63	K	15	offen	offen	offen	ADV	ADJD	{‘Degree’: ‘Pos’}		oc
63	K	16	und	und	und	CCONJ	KON			cd
63	K	17	da	da	da	ADV	ADV			mo
63	K	18	ist	ist	sein	VERB	VAFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘’}	v	cj
63	K	19	so	so	so	ADV	ADV			mo
63	K	20	bile	viele	vieler	PRON	PIAT	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Plur’, ‘PronType’: ‘Ind’}		sb
63	K	21	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}	sb	pd
63	K	22	und	und	und	CCONJ	KON			cd
63	K	23	kann	kann	können	AUX	VMFIN	{‘Mood’: ‘Ind’, ‘Number’: ‘Sing’, ‘Person’: ‘3’, ‘Tense’: ‘Pres’, ‘VerbForm’: ‘Fin’}	v	cj
63	K	24	dach	Dach	Dach	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Neut’, ‘Number’: ‘Sing’}		mo
63	K	25	da	da	da	ADV	ADV			mnr
63	K	26	badi	Badi	Badi	NOUN	NN	{‘Case’: ‘Nom’, ‘Gender’: ‘Fem’, ‘Number’: ‘Sing’}		mo
63	K	27	lauffen	laufen	laufen	VERB	VVINF	{‘VerbForm’: ‘Inf’}		oc

Table 20: Swiss Standard German language sample annotated. Words like “abeR” with R in upper case indicate the emphasizing tone.

Resource-Efficient LLMs for Depression Symptoms Screening: Performance and Limitations in Zero Shot Setting

Muhammad Rizwan, Jure Demšar
Faculty of Computer and Information Science
University of Ljubljana, Večna pot 113, Ljubljana 1000, Slovenia
{muhammad.rizwan, jure.demsar}@fri.uni-lj.si

Abstract

Depression is the leading cause of global disability and early detection is crucial for effective intervention. Recent advances in large language models (LLMs) offer potential for analyzing text to identify depression symptoms. This work investigates the zero-shot capability of LLMs to recognize nine DSM5 depression symptoms from short-text inputs. We evaluated eight open LLMs with model sizes ranging from 1.5B to 14B parameters using a clinically annotated dataset and assessed both overall agreement and symptom-level performance. Results indicate that while smaller models exhibit limited clinical accuracy, the Qwen 2.5-7B model achieves substantial performance with a Cohen's Kappa of 0.603 and a Macro F1 score of 0.648. Notably, a performance plateau between the 7B and 14B Qwen variants suggests that model scaling alone does not guarantee improved symptom-level classification, establishing Qwen 2.5-7B as a resource-efficient model. Further analysis of the best-performing model revealed strengths in identifying salient symptoms like suicidal thoughts, but limitations in recognizing core symptoms such as depressed mood and anhedonia. Misclassification analysis reveals that the model frequently misclassifies posts expressing 'depressed mood' as 'no symptom' or vice versa, often overlooking indicators of irritability or social withdrawal. These findings suggest that resource-efficient LLMs can support preliminary symptom screening in zero shot settings, but there is risk of overlooking clinically important symptoms without fine-tuning.

Keywords: DSM-5, Depression Symtpoms, Large Language Models, Mental Health

1. Introduction

According to the World Health Organization (WHO), depression is one of the leading causes of global disability, affecting an estimated 5.7% of the adult population worldwide. Furthermore, it contributes to approximately 727,000 suicide deaths annually[1]. In Europe, mental health conditions, including chronic depression, affect about 7% of the population. In response to this growing burden, 31 countries in the WHO European Region have committed to integrating mental health into all areas of public policy and prioritizing it within national health agendas[2].

Despite the urgent need for large-scale mental health assessment, traditional clinical evaluation remains resource-intensive and difficult to scale. In recent years, social media platforms have emerged as valuable sources of user-generated content that reflects individuals' emotional states and lived experiences. This development has motivated a growing body of research in natural language processing (NLP) aimed at detecting mental health signals from textual data.

Early computational studies demonstrated correlations between linguistic features and depression-related behaviors on platforms such as Twitter and Reddit, using lexicons, topic models, and traditional machine learning classifiers (Liu et al., 2022; De Choudhury et al., 2013; Coppersmith et al., 2014). However, much of this work framed depression detection as a binary or multi-class diagnosis prediction task, frequently relying on self-disclosed diagnoses as ground truth labels. Clinical researchers have criticized such labeling strategies for their noise, demographic bias, and limited clinical validity (Ernala et al., 2019).

To address these concerns, more recent research has shifted from disorder-level classification to symptom-level modeling. Instead of predicting a diagnosis, these approaches aim to detect individual psychological or behavioral indicators corresponding to specific diagnostic criteria. Studies have explored mapping social media language to DSM-5 depression symptoms such as anhedonia, sleep disturbances, and feelings of worthlessness (Manikonda and De Choudhury, 2017; Chancellor et al., 2019). At the same time, there is an increasing trend toward constructing datasets annotated by domain experts and developing standardized evaluation protocols for depression and suicide risk detection (Zhang et al., 2021; Coppersmith et al., 2014). These developments provide an opportunity to evaluate modern NLP systems against clinically

[1] https://www.who.int/en/news-room/fact-sheets/detail/depression

[2] https://www.who.int/europe/news/item/16-06-2025-with-17–of-people-in-the-region-living-with-a-mental-health-condition–31-countries-commit-to-integrating-mental-health-into-all-policies

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 74–81
12 May 2026.

grounded, human-annotated benchmarks.

The emergence of large language models (LLMs) has introduced new possibilities for mental health text analysis. Instruction-tuned and conversational LLMs demonstrate strong generalization capabilities across diverse tasks without task-specific training, enabling zero-shot and few-shot learning paradigms. Recent studies have explored their application to mental health-related tasks, including depression detection, suicide risk assessment, and emotional support generation (Yang et al., 2023; Lan et al., 2025; Jin et al., 2025; Omar et al., 2024). Zero-shot classification using natural language prompts has become a prominent approach for evaluating LLM generalization, where task labels are framed as natural language descriptions to leverage pretrained knowledge without explicit supervision (Zhao et al., 2023; Kojima et al., 2022). In mental health contexts, prompt-based methods offer the advantage of explicitly incorporating clinical definitions, potentially improving alignment with expert annotations.

Nevertheless, the reliability of LLMs for clinically grounded, symptom-level mental health analysis remains insufficiently studied. In particular, relatively few works systematically examine zero-shot performance of resource-efficient, open LLMs for detecting DSM-5 depression symptoms. The extent to which such models can replicate expert-level symptom identification without fine-tuning remains unclear.

In this work, we investigate whether moderately sized, general-purpose open LLMs can identify depression-related symptoms in short texts under zero-shot conditions. We evaluate their agreement with DSM-5 symptom annotations provided by licensed psychologists, focusing particularly on clinically critical symptoms such as suicidal thoughts, worthlessness, and anhedonia. Through a controlled zero-shot evaluation, we aim to assess both the feasibility and the limitations of LLM-based symptom recognition systems and to identify common misclassification patterns across DSM-5 symptom categories[3].

2. Methods

This section details the methodology employed in our study. We first describe the preparation of the ReDSM5 dataset used for training and evaluation. Then, we describe the experimental setup, including the LLMs used in experiments and the prompt-based inference procedure used to assess their zero-shot depression symptom classification capabilities.

[3]Code: https://github.com/rizwan2phd/zeroshot-depression-symptoms-screening-llms

Symptom	Sentence Count
Depressed mood	326
Worthlessness	284
Suicidal thoughts	175
Fatigue	111
Anhedonia	106
Sleep issues	104
Cognitive issues	53
Appetite change	45
Psychomotor	32
None (control)	374

Table 1: **Distribution of DSM-5 Depression Symptoms in the ReDSM5 Dataset**. This table displays the number of sentences labeled by a licensed psychologist as containing each of the listed DSM-5 depression symptoms, alongside a 'None' control class.

2.1. Dataset Preparation

In this study we used the ReDSM5 dataset (Bao et al., 2025), a clinically labeled Reddit corpus curated according to DSM5 depression standards (Tolentino and Schmidt, 2018), (Information Retrieval Lab, University of A Coruña, 2025). The dataset contains Reddit sentences that were reconstructed from an earlier paragraph-level corpus and re-annotated by a licensed clinical psychologist. Every sentence is labeled for the presence or absence of the nine DSM5 depression symptoms: depressed mood, worthlessness, suicidal thoughts, anhedonia, fatigue, sleep issues, cognitive issues, appetite change and psychomotor. In addition to binary symptom labels, the annotator also provided clinical rationales.

The instances with multiple symptom labels represent a relatively small proportion of the overall dataset. To simplify the task for our model and facilitate clear interpretation of results, we removed these multi-label instances, reducing the dataset size by 129 instances. The dataset also contains sentences with the absence of symptoms, we used those as our control class (symptom = none).

This resulted in a refined dataset of 1,610 single label instances encompassing ten unique categories within the DSM5 symptoms, nine original symptom labels and a new `None` category representing the absence of depressive symptoms (control class). The final distribution of the data can be seen in Table 1.

2.2. Experiment Setup

To evaluate the zero-shot depression symptom classification capabilities of LLMs, we employed a fixed prompt-based inference setup using the eight open models with sizes ranging from 1.5B to 14B param-

eters and model families (`Llama`, `Mistral`, `Qwen`). List of all models with parameter size and corresponding performance can be seen in Table 2. The Qwen family (1.5B, 3B, 7B and 14B parameters) is composed of models trained on large scale multilingual corpora with strong coverage of English and Chinese language. These models are instruction tuned to improve reasoning. The latest LLaMA 3.2 family (1B and 3B parameters) represent compact and efficient instruction tuned models, while the Mistral family (7B and 12B parameters) represent instruction tuned models employ optimized Transformer architectures. Details of the models evaluated and their performance are presented in Table 2.

Through the prompt, we asked the model to perform the task of an expert psychologist and label the input text exactly one of the DSM5 major depressive disorder symptoms i.e. depressed mood, worthlessness, suicidal thoughts, anhedonia, fatigue, sleep issues, cognitive issues, appetite change and psychomotor or assign the none label if the text does not fit any of these labels. The exact prompt was as follows:

> You are an expert clinical psychologist trained in DSM5 diagnostic criteria for major depressive disorder. Analyze the text carefully for indicators of depressive symptoms. If the text describes a depressive symptom, classify it into one category. If the text shows NO depressive symptoms (e.g., normal mood, neutral content, unrelated topics), classify it as NONE.
> TASK: Classify the following text into EXACTLY ONE category from the list below. Categories: DEPRESSED MOOD, WORTHLESSNESS, ANHEDONIA, SUICIDAL THOUGHTS, APPETITE CHANGE, SLEEP ISSUES, FATIGUE, COGNITIVE ISSUES, PSYCHOMOTOR.
> Output ONLY the category label, no explanations, no punctuation, no additional text. Choose NONE if the text shows NO depressive symptoms. Your response must be a single word matching one category exactly.

We constrained the LLM output to return only the category label to ensure consistency and eliminate post-processing ambiguity. The `max_tokens` parameter was set to 15 to further enforce single-label output. Any response failing to produce an exact category label from the predefined list, or containing extraneous text, was considered an invalid prediction and excluded from the evaluation. The last column of Table 2 shows the number of valid predictions per model. All experiments were conducted in a zero-shot setting, meaning no task-specific fine-tuning or in-context examples were provided. To ensure deterministic outputs and reproducibility, the temperature parameter was fixed at 0 for all model inferences.

This setup allows us to assess whether computationally efficient models can effectively generalize to depression symptoms identification under strict zero-shot condition.

3. Results and Discussion

This section discusses the performance of LLMs for zero-shot classification of DSM5 depression symptoms. We first present a comprehensive analysis of model agreement and overall performance, and then analyze the symptom-level insights and misclassification patterns to identify both strengths and limitations of our best performing model.

3.1. Human Agreement and Performance Analysis of DSM5 Depression Symptoms

Our zero-shot evaluation of the models for DSM5 depression symptom classification shows clear performance differences across model sizes and architectures. Models scores can be seen in Table 2 and visualized in Figures **??**. Given the class imbalance in the dataset, we rely on Cohen's Kappa (McHugh, 2012) and Macro F1 as evaluation metrics.

The Macro F1 score is calculated as:

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i \tag{1}$$

where N is the number of classes and F1_i is the F1-score for class i. The class-wise F1-score is calculated as:

$$\text{F1}_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{2}$$

Thus, the Macro-F1 score in Equation 1 is computed as the average of the class-wise F1-scores defined in Equation 2.

Cohen's Kappa (κ) is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where p_o is the observed agreement and p_e is the expected agreement by chance, as shown in Equation 3.

Overall, extremely small models struggle to perform meaningful clinical agreement in a zero-shot setting. *Llama-3.2-1B-Instruct*, for instance, exhibits small agreement (Cohen's $\kappa = 0.060$) and a very low Macro F1 score (0.089). This indicates that models at this scale are incapable of capturing DSM5 symptom distinctions without additional

Model	Parameters	Cohen Kappa	Macro F1	Valid Predictions
Llama-3.2-1B-Instruct	1B	0.060	0.089	1608
Llama-3.2-3B-Instruct	3B	0.412	0.410	1603
Mistral-7B-Instruct-v0.3	7B	0.511	0.587	1523
Mistral-Nemo-Instruct-2407	12B	0.550	0.607	1601
Qwen2.5-1.5B-Instruct	1.5B	0.343	0.349	1572
Qwen2.5-3B-Instruct	3B	0.491	0.532	1602
Qwen2.5-7B-Instruct	**7B**	**0.603**	**0.648**	1605
Qwen2.5-14B-Instruct	14B	0.600	0.651	1537

Table 2: **Zero-Shot Classification Results of LLMs for DSM-5 Depression Symptoms**. Performance metrics (Cohen's Kappa, Macro F1) for several LLMs are shown, evaluated on the ReDSM5 dataset in zero shot settings. This highlights the potential for direct application of LLMs to mental health assessment. Valid predictions represent sentences where the LLM correctly outputs the exact symptom category. Qwen 2.5-7B-Instruct achieved the best overall performance.

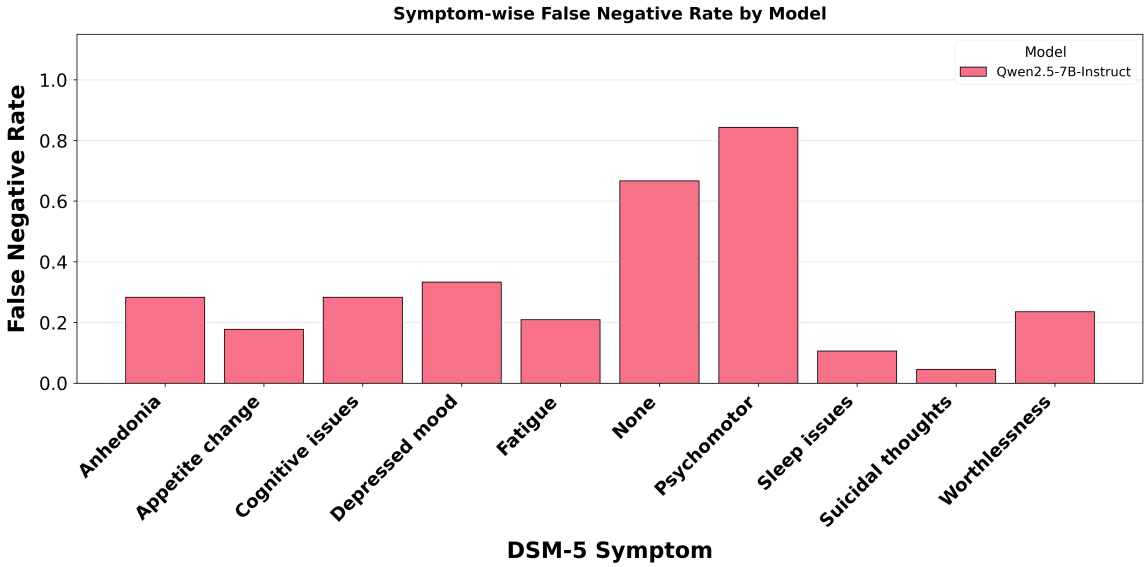


Figure 1: **Symptom-wise false negative rate of the Qwen2.5-7B-Instruct model.** Higher rates indicate a tendency to miss identifying specific symptoms.

human guidance. Performance improves substantially as model size increases into the 3B–7B range, which is the minimum threshold necessary for reliable zero-shot symptom classification. In particular, the *Qwen 2.5* series consistently outperforms size-matched alternatives, which means Qwen architecture and training better suited for depression symptom classification. Among all evaluated models, *Qwen 2.5-7B* emerges as the strongest. It achieves the highest overall agreement with DSM5 labels, with a fair Cohen's κ of 0.603 and a Macro F1 score of 0.648, marginally surpassing both the larger *Qwen 2.5-14B* and *Mistral-Nemo-12B* models. It shows that *Qwen 2.5-7B* offers superior parameter efficiency and more effective zero-shot clinical reasoning.

Interestingly, the performance plateau observed between the 7B and 14B Qwen variants implies that scaling alone does not improve symptom-level classification in the absence of fine-tuning or few-shot prompting. This shows that representational quality of model is more influential than only model size for the zero-shot settings. Based on these findings, we select *Qwen 2.5-7B* as the best performing model for symptom level analyses in upcoming sections.

3.2. Symptom Level Performance

As shown in Figure 3, there is a clear performance plateau between Qwen 2.5-7B and Qwen 2.5-14B, with nearly identical F1 scores across most symptom categories. Although the 14B model demonstrates marginal improvements on a few symptoms (e.g., anhedonia and appetite change), these gains are small and inconsistent. Given this, the

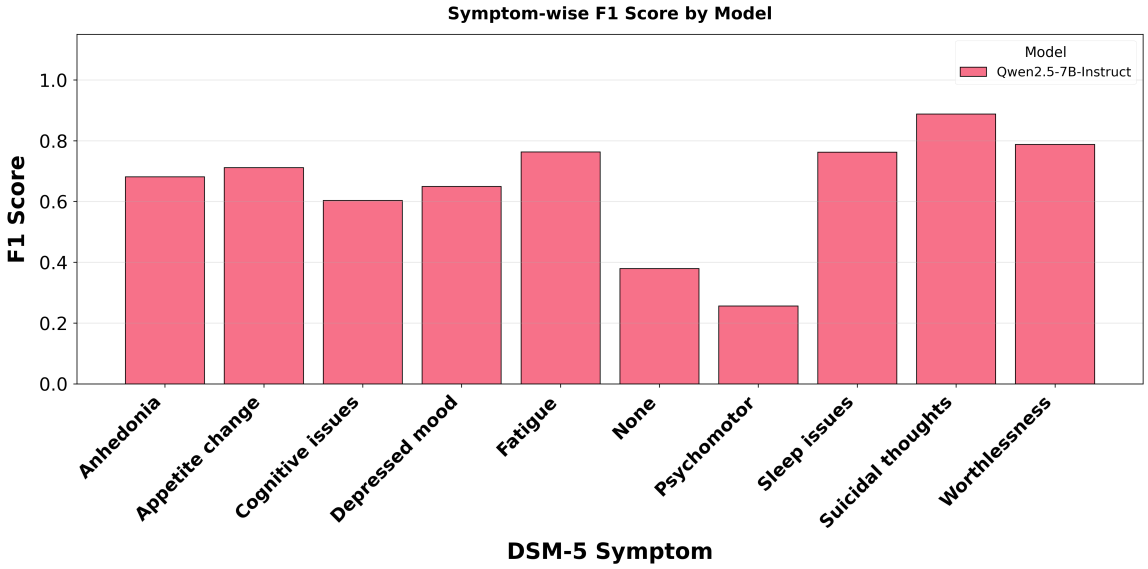


Figure 2: **Symptom-specific F1 scores for Qwen2.5-7B-Instruct.** The figure presents F1 scores across all DSM-5 symptoms, including the None category, illustrating the model's performance in symptom-level detection under zero-shot settings.

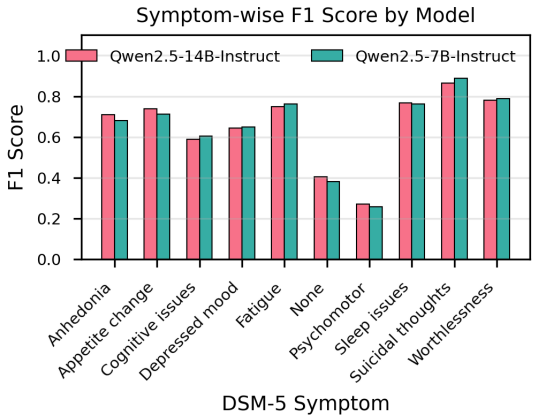


Figure 3: **Symptom-wise F1 comparison.** Comparison of F1 scores across DSM-5 symptoms for *Qwen2.5-7B-Instruct* and *Qwen2.5-14B-Instruct*, showing minimal performance differences between models.

efficiency–performance trade-off favors Qwen 2.5-7B, which achieves comparable results with lower computational cost. Consequently, the 7B model represents a practical choice for deployment and a strong candidate for further fine-tuning to achieve additional performance gains.

The symptom-wise evaluation of our best-performing model, Qwen 2.5-7B, on DSM-5 depression criteria reveals substantial variability across symptoms in terms of F1 score and false-negative rate (FNR). We analyze both the model's strengths and its clinically relevant limitations. The model achieves strong performance on linguistically explicit and high-salience symptoms such as suicidal thoughts (F1 = 0.888), worthlessness (F1 = 0.788), and sleep issues and fatigue (F1 $\approx$ 0.76). Notably, the high recall for suicidal thoughts suggests potential utility in high-sensitivity screening settings. Symptom-wise F1 and FNR scores are shown in Figure 1 and Figure 2, respectively.

In contrast, the model exhibits only moderate performance on core symptoms, including depressed mood (FNR = 0.333) and anhedonia (FNR = 0.283). These elevated false-negative rates are clinically concerning, as prior work (Loas et al., 2018; Auerbach et al., 2022; Gillissie et al., 2023) highlights a strong association between anhedonia and suicidal thoughts. Moreover, suicidal thoughts, depressed mood, and anhedonia are among the most critical symptoms for assessing severe depression risk (Zimmerman et al., 2018). The model performs worst on psychomotor symptoms, with a very high false-negative rate (FNR = 0.884), indicating frequent failure to detect their presence.

Additionally, the *none* category, representing the absence of depressive symptoms, shows poor performance (F1 = 0.379). This indicates a systematic bias toward predicting symptom presence even when no symptoms are present, suggesting that the model may misinterpret neutral or mildly concerning language as clinically relevant, potentially leading to unnecessary alerts in real-world applications.

True Label	Predicted Label	Counts	Sample Sentences (Separated by semi colon)
Depressed Mood	None	72	I know that life is pretty tough and I try to work and safe to have a more secure future; i am angry all the time; Also I have lived recklessly because I thought I didn't had much future; But I never cry, no matter how sad; Where I am neither very happy or sad; I get irritated easily; I'm so irritable all the time.
None	Depressed Mood	66	I feel sad from time to time; I cried so much; Makes me sad ,though; Then I get sad; I have watched it 3 times and I still cry every time; Iḿ sad and I fear marriage now; I wake up feeling depressed, upset, and anxious; But I never miss my sadness because that's depression and it could get so bad;
None	Sleep Issues	41	Now, i never feel rested in the morning even if i slept 8-9 hours during the night, and it takes a long time to fall asleep; I usually just get six hours of poor sleep every night; I tend to fell asleep and laid down as soon as I get home; Some nights I would get no sleep at all.

Table 3: : **Misclassification analysis of Qwen 2.5-7B predictions.** The table presents the most frequent misclassifications, including the true label, predicted label, the number of occurrences, and representative example sentences.

3.3. Misclassification Pattern Analysis

In this part of research work, we present results from a detailed exploration of the most common misclassification produced by Qwen 2.5 7B during zero-shot classification of DSM5 depressive symptoms, see Table 3. This is crucial for understanding the limitations of the model when applied to clinical data and for guiding future model improvement.

The most frequent misclassification pattern is to classify depressed mood as none(absence of a symptom). The model struggles to understand that depression does not always present as sadness alone. In several cases, explicit denials of sadness such as *"I wasn't sad or crying anymore"*, *"I don't feel overwhelming sadness"* or *"But I never cry"* lead the model to assume that depression is not present.

Instead of recognizing irritability and anger as possible expressions of a depressed mood, it often treats them as separate, unrelated emotions. Statements such as *"I'm becoming extremely irritable"*, *"constantly mad at little things"* and *"get really irritable"* are frequently misclassified. This suggests a narrow interpretation of depressive symptoms.

Similarly, the model often overlooks signs of social withdrawal. Expressions like *"I had no friends"* or *"Feel alone, that no one has my back"* point to isolation, yet these cues are not consistently recognized.

The second misclassification pattern is the opposite of the first one – classifying sentences without a symptom(the none label) as sentences with a depressed mood symptom. It appears to rely heavily on keywords such as -sad- or -cry- without adequately distinguishing between short-term emotional responses and the persistent low mood. This pattern is especially evident in sentences describing sadness triggered by specific events or ordinary emotional experiences, such as *"I feel sad from time to time"*, *"I cried so much"*, *"Makes me sad though"*, *"Then I get sad"* or *"I have watched it three times and I still cry every time."*

Another prominent pattern among the misclassified instances is the use of quantified sleep-duration constructions without explicit negative qualifiers, typically when explicitly using the keyword sleep along with number of hours. Examples include statements such as *"I normally sleep for 7–8 hours"*, *"I slept 15 hours yesterday"* and *"I sleep 12 hours without waking up"*. Although these sentences describe excessive sleep in context, they are linguistically framed as neutral behavioral reports rather than explicit complaints. The absence of strong distress markers (e.g., insomnia, can't sleep, exhausted) may cause models to interpret them as routine or lifestyle descriptions, leading to misclassification into the control class despite their association with sleep disturbance and depressive symptoms.

4. Conclusion

This study demonstrates the potential of zero-shot, open LLMs for identifying depression symptoms from short text. Our results show notable perfor-

mance differences across model sizes and architectures. We tested several models of different sizes that belongs to different model families. In our case, Qwen 2.5-7B shows the best results for identifying DSM5 depression symptoms. In particular, it achieved fair agreement with DSM5 criteria (Cohen's $\kappa = 0.603$), highlighting its capacity to approximate clinically relevant symptom classification. Furthermore, a symptom-level analysis further revealed that the model performs well in detecting more explicit indicators, such as suicida thoughts, while encountering difficulties to identify depressed mood and anhedonia. Additionally, we observed a systematic bias toward predicting the presence of symptoms even in neutral text, underscoring the need for improved calibration.

Our findings clarify where the model performs reliably and where it falls short, providing direction for targeted refinements. Especially to reduce critical false negatives in high-risk symptom categories. Overall, Qwen 2.5-7B shows promise as a supportive tool for preliminary screening and automated symptom detection. However, we believe that such approaches are not yet ready and should not be used as a substitute for professional clinical assessment.

5. Limitations and Future Work

This study has limitations that should be acknowledged when interpreting the results. First, our evaluation focused on zero-shot performance with a fixed prompt. While this approach was chosen to assess the models' general suitability for initial depressive symptom screening, performance may vary with different prompt formulations. Second, we utilized a single, clinically-annotated dataset (ReDSM5) derived from Reddit posts. This introduces potential bias as the data may not fully represent the broader population of individuals experiencing depression. Furthermore, to simplify analysis, we focused solely on single-label instances, neglecting the common occurrence of co-occurring depressive symptoms represented in the dataset's limited multi-label examples.

Future research address the limitations by exploring few-shot learning approaches and investigating performance on multi-label instances. Specifically, targeted experiments with few-shot learning could reveal whether smaller models within the Qwen family can achieve improved performance in identifying depressive symptoms.

6. Ethics Statement

This study uses the ReDSM5 dataset, which is available under restricted access subject to specific terms and conditions. The data are anonymized and used exclusively for research purposes, with no attempt to identify or re-identify individuals. Given the sensitive nature of mental health content, the proposed study is not intended for clinical diagnosis or deployment without expert supervision.

7. Funding

This publication has received funding from the European Union's Horizon Europe research and innovation program under the Marie Sklodowska-Curie COFUND Postdoctoral Programme grant agreement No.101081355- SMASH and by the Republic of Slovenia and the European Union from the European Regional Development Fund.

8. Disclaimer

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Exacutive Agency. Neither the European Union nor the granting authority can be held responsible for them.

9. References

Randy P Auerbach, David Pagliaccio, and Jaclyn S Kirshenbaum. 2022. Anhedonia and suicide. *Anhedonia: Preclinical, translational, and clinical integration*, pages 443–464.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2025. Redsm5: A reddit dataset for dsm-5 depression detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6323–6327.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology From linguistic signal to clinical reality*, pages 51–60.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16.

Emily S Gillissie, Gia Han Le, Taeho Greg Rhee, Bing Cao, Joshua D Rosenblat, Rodrigo B Mansur, Roger C Ho, and Roger S McIntyre. 2023. Evaluating anhedonia as a risk factor in suicidality: a meta-analysis. *Journal of psychiatric research*, 158:209–215.

Yu Jin, Jiayi Liu, Pan Li, Baosen Wang, Yangxinyu Yan, Huilin Zhang, Chenhao Ni, Jing Wang, Yi Li, Yajun Bu, et al. 2025. The applications of large language models in mental health: scoping review. *Journal of Medical Internet Research*, 27(1):e69284.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. 2025. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2155–2171.

Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, Jing Guo, et al. 2022. Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Mental Health*, 9(3):e27244.

Gwenolé Loas, Guillaume Lefebvre, Marianne Rotsaert, and Yvon Englert. 2018. Relationships between anhedonia, suicidal ideation and suicide attempts in a large sample of physicians. *PloS one*, 13(3):e0193619.

Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and understanding visual attributes of mental health disclosures in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 170–181.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Mahmud Omar, Shelly Soffer, Alexander W Charney, Isotta Landi, Girish N Nadkarni, and Eyal Klang. 2024. Applications of large language models in psychiatry: a systematic review. *Frontiers in psychiatry*, 15:1422807.

Julio C Tolentino and Sergio L Schmidt. 2018. Dsm-5 criteria and depression severity: implications for clinical practice. *Frontiers in psychiatry*, 9:450.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Yipeng Zhang, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, and Jiebo Luo. 2021. Monitoring depression trends on twitter during the covid-19 pandemic: observational study. *JMIR infodemiology*, 1(1):e26769.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2023. Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 15590–15606.

Mark Zimmerman, Caroline Balling, Iwona Chelminski, and Kristy Dalrymple. 2018. Understanding the severity of depression: which symptoms of depression are the best indicators of depression severity? *Comprehensive psychiatry*, 87:84–88.

10. Language Resource References

Information Retrieval Lab, University of A Coruña. 2025. *ReDSM5: A Reddit Dataset for DSM-5 Depression Detection*. Information Retrieval Lab (IRLab), Universidade da Coruña. Hugging Face Datasets, Mental Health and Clinical NLP Resources, 1.0. PID https://huggingface.co/datasets/irlab-udc/redsm5. Reddit corpus annotated with DSM-5 depressive symptoms. Gated access via Hugging Face. Dataset described in Bao et al., accepted at CIKM 2025.

CNSocialDepress: A Chinese Social Media Dataset for Depression Risk Detection and Structured Analysis

Jinyuan XU[1*], Tian LAN[2*], Xintao YU[3], Xue HE[3,4], Hezhi ZHANG[5], Ying WANG[6]
Pierre Magistry[1], Mathieu Valette[1], Lei LI[7†]
[1]Ertim Inalco, [2]Milkuya Studio, [3]Sorbonne Université, [4]IRD Lab,
[5]Faculty of Psychology, Peking University,
[6]Faculty of Psychology and Cognitive Science, Beijing Normal University,
[7]Beijing Institute of Technology

Abstract

Depression is a pressing global public health issue, yet publicly available Chinese-language resources for depression risk detection remain scarce and largely focus on binary classification. To address this limitation, we release **CNSocialDepress**, a benchmark dataset for depression risk detection on Chinese social media. The dataset contains 44,178 posts from 233 users; psychological experts annotated 10,306 depression-related segments. CNSocialDepress provides binary risk labels along with structured, multidimensional psychological attributes, enabling interpretable and fine-grained analyses of depressive signals. Experimental results demonstrate the dataset's utility across a range of NLP tasks, including structured psychological profiling and fine-tuning large language models for depression detection. Comprehensive evaluations highlight the dataset's effectiveness and practical value for depression risk identification and psychological analysis, thereby providing insights for mental health applications tailored to Chinese-speaking populations.

Keywords: Depression Detection, Chinese Social Media, Benchmark Dataset, Mental Health

1. Introduction

Depressive disorders are among the most common mental health conditions worldwide and are characterized by persistent low mood or loss of interest in daily activities. According to a 2023 report[1] by the World Health Organization (WHO), approximately 280 million people worldwide live with depression. In China, a 2024 survey by the Chinese Center for Disease Control and Prevention (Wang et al., 2024a) estimates that around 95 million individuals are affected by depression. Of the approximately 280,000 suicides reported annually in China, 40% are linked to depressive disorders. Moreover, research (Wang et al., 2024a) shows that depression is strongly associated with both suicidal behavior and non-suicidal self-injury.

Motivated by the urgent need to detect depression, researchers increasingly apply machine learning (ML) and natural language processing (NLP) methods to automatically assess depression risk (Squires et al., 2023; Hasib et al., 2023; Aleem et al., 2022; Liu et al., 2024; Shi et al., 2024; Jia and Li, 2024). While early efforts have shown promising results, they remain constrained by the limitations of existing datasets. Traditional depression detection studies often rely on clinical data (Bittar et al., 2019; Fernandes et al., 2018) or on transcripts from medical or psychological interviews (Shen et al., 2022; Li et al., 2022b), which are expensive to collect, limited in size and diversity, and may not reflect the informal, emotionally nuanced expressions typical of real-world online environments.

To overcome these limitations, recent research has shifted toward leveraging user-generated content on social media platforms (Bucur et al., 2025; Harrigian et al., 2021; Li et al., 2025d; Jiang et al., 2024; Cai et al., 2025a), which provides rich linguistic signals and is more readily available. Such data are also more easily anonymized, thereby alleviating some privacy concerns. Beyond these methodological advantages, social media has become an important channel for emotional expression, especially in East Asian contexts (Zhou et al., 2023; Yang and Li, 2009; Zhou et al., 2024; Yang et al., 2025), where personal emotions may be expressed more freely online than in face-to-face settings. As a result, social media can provide richer and more authentic linguistic signals for depression analysis.

However, most existing datasets focus solely on classification with binary or multi-class labels and lack structured psychological insights or professional validation. Moreover, generative models are increasingly used in mental health applications (Hu et al., 2024; Xu et al., 2024; Yang et al., 2024b; Lai et al., 2023; He and Qu, 2025a; Gu et al., 2025; Cai et al., 2025b; Xu et al., 2025). Researchers in psychology and computational linguistics emphasize the need for datasets that combine risk labels with structured analyses, such as user profiles or

*These authors contributed equally to this work.
†Corresponding Author.
[1]https://www.who.int/news-room/fact-sheets/detail/depression

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 82–105
12 May 2026.

survey-based explanations. Such datasets better match real-world settings. Multifaceted insights can support clinical decision-making and downstream intervention.

Some recent efforts have introduced summarization-style datasets for depression. They leverage information extraction and automated summarization techniques (Sotudeh et al., 2022, 2021; He and Qu, 2025b; Shi et al.). However, these datasets often rely on model-generated content. The content is typically validated using automatic metrics such as ROUGE (Lin, 2004). They also lack annotation by mental health professionals, which raises concerns about domain-specific reliability and accuracy.

To address these challenges, we introduce **CNSocialDepress** (**CNSD**), a publicly available Chinese-language dataset for depression risk detection that can be applied to the development of early-stage depression detection tools. It pairs binary risk labels with structured psychological analyses. All annotations are drafted and validated by certified mental health professionals, ensuring domain relevance and annotation quality. **CNSD** supports multiple task paradigms. These include binary classification, structured analysis generation, summarization, and fine-tuning large language models (LLMs) for psychological reasoning.

Our contributions are fourfold:

- We release **CNSD**, a high-quality Chinese dataset for depression risk detection that combines binary labels with structured, interpretable psychological analyses.
- We propose an expert-in-the-loop annotation protocol with structured templates and quality control.
- We benchmark **CNSD** across diverse task settings, including classification, structured analysis generation, summarization, and LLM fine-tuning for psychological reasoning.
- We present and validate a pipeline for generating structured psychological analyses for depression risk (Section 4).

2. Related Work

2.1. Existing Datasets for Depression Detection

Current datasets for depression detection mainly come from English-language social media platforms such as Twitter, Reddit, Facebook, and Instagram (Shen et al., 2017; Parapar et al., 2022; Zhang et al., 2021; Raihan et al., 2024; Li et al., 2025c; Jiang et al., 2025). Common annotation strategies include identifying self-disclosure (Yates et al., 2017; Bathina et al., 2021; Islam et al., 2018), manual coding by clinical raters (Almouzini et al., 2019; Yazdavar et al., 2020; Alhamed et al., 2024), and symptom mapping (Seabrook et al., 2018; Aldarwish and Ahmad, 2017; Zhang et al., 2022; Li et al., 2025a). Symptom definitions are often based on DSM-5 criteria (Association, 2013) and questionnaire instruments such as the PHQ-9 (Kroenke et al., 2001).

Most datasets use binary labels. Others incorporate severity scales or symptom-specific tags (Zhang et al., 2022; Mowery et al., 2017; Guan et al., 2025). Multilingual resources also exist for depression detection. They cover languages such as Spanish (Romero et al., 2024), Arabic (Maghraby and Ali, 2022), Russian (Stankevich et al., 2020), Portuguese (Santos et al., 2024), Japanese (Yuka Niimi, 2021), and Thai (Hämäläinen et al., 2021). For Chinese, Sina Weibo is the most common data source (Li et al., 2020; Shen et al., 2018; Li et al., 2023; Guo et al., 2023; Yang et al., 2021; Li et al., 2025b). Datasets such as WU3D (Wang et al., 2020a) and SWDD (Cai et al., 2023) are widely used.

2.2. Methodological Advancements

Early work relied on statistical representations such as TF–IDF and hand-crafted features (Yang et al., 2020; Li et al., 2022a). These features were typically paired with traditional machine learning classifiers (Cortes, 1995; Breiman, 2001; McCallum et al., 1998; Dreiseitl and Ohno-Machado, 2002; He et al., 2026). With the rise of neural approaches (Schmidhuber, 2015), representation learning became central to depression detection. Embedding methods such as Word2Vec (Mikolov et al., 2013) and neural architectures including LSTMs (Sak et al., 2014), CNNs (Kim, 2014), Transformers (Vaswani, 2017), and BERT (Devlin et al., 2019) improved performance.

Recently, large language models (LLMs) have been explored for depression detection (Lan et al., 2024; Hu et al., 2024; Xin Yan, 2023; Lai et al., 2023; Li et al., 2026; He and Qu, 2025a; Shi et al., 2025). They can produce predictions and generate natural-language rationales or structured analyses, which supports interpretability. LLM-based approaches have also been used to improve explainability in mental health analysis (Wang et al., 2024c; Yang et al., 2024b; Xu et al., 2024; Yang et al., 2023; Hu et al., 2024; Gu et al., 2025).

2.3. Summary and Gaps

Despite these efforts, few datasets include structured psychological analyses, especially for non-English social media. Furthermore, available resources often lack annotation by mental health

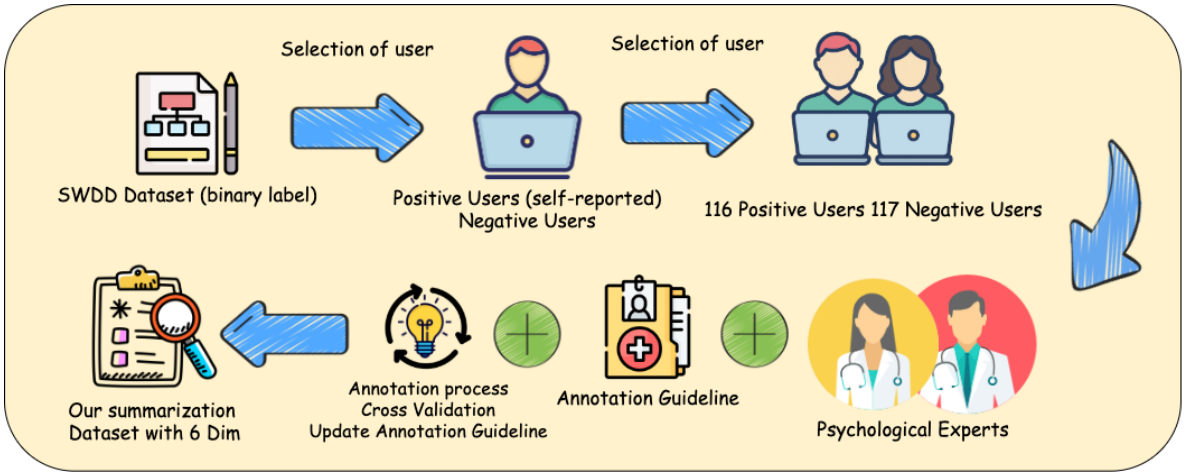


Figure 1: Dataset construction process. We sample 116 positive (depressed) users and 117 negative users from SWDD for expert annotation. Psychologists draft an initial guideline based on DSM-5 and corpus statistics, and iteratively refine it through random cross-validation. The final gold-standard data include a six-dimensional structured analysis summary for each user.

professionals, which limits practical utility. This study addresses these limitations by presenting an expert-validated Chinese-language dataset with detailed structured annotations. The dataset provides richer resources for depression risk detection and analysis.

3. Construction of Dataset

Our raw data come from the SWDD dataset (Cai et al., 2023), a user-level corpus collected from Sina Weibo, one of the largest Chinese-language social media platforms. The dataset includes two user groups: depressed and non-depressed. Each user has dozens to hundreds of Weibo posts. In addition to binary user labels, SWDD provides expert-annotated user-level depression features based on DSM-5 criteria[2] (Association, 2013).

From SWDD, we select 116 self-reported depressed (positive) users who disclosed a clinical diagnosis of depression in their posts, and 117 non-depressed (negative) users as candidates. Each selected user has at least 60 posts. The total length of a user's posts is at least 3,000 tokens, computed using the Qwen2.5 tokenizer.[3]

To guide annotation, a team of psychology experts drafted an initial guideline based on DSM-5 criteria, the PHQ-9 (Kroenke et al., 2001), and prior work on linguistic markers of depression in online text (Mothe et al., 2022).

We define two levels of criteria. Primary criteria cover objective statements with higher diagnostic specificity, such as clinical symptoms, medical records, and explicit self-reports. Secondary criteria capture subjective linguistic or emotional expressions with lower specificity, such as negative phrasing and emotional intensity. This hierarchy follows clinical principles that prioritize objective symptoms over subjective perceptions. The guidelines cover six dimensions:

- **Dimension 1: Depressive psychological state (primary).** This dimension reflects loss of self-worth, overwhelming guilt, and suicidal ideation. Representative spans include *inferiority*, *apology*, *death*, *self-harm*, and *suicide*.
- **Dimension 2: Medical expressions related to depression (primary).** This dimension covers mentions of medication use and clinical diagnoses, such as *taking medication*, *venlafaxine*, *side effects*, *depression*, *anxiety disorder*, *hospital*, and *doctor*.
- **Dimension 3: Clinical symptoms related to depression (primary).** This dimension captures physiological or somatic changes, including appetite or weight changes, sleep disturbance, fatigue, and physical pain. Representative spans include *appetite*, *insomnia*, *sleeping pills*, *nightmares*, *tiredness*, *headache*, and *stomachache*.
- **Dimension 4: Negative emotions (secondary).** This dimension covers adverse emotional states such as *sadness*, *grief*, *anxiety*, *loneliness*, and *despair*.
- **Dimension 5: Potential external causes of depression (secondary).** This dimension captures possible triggers such as intimate relationship issues, family conflict, major social

[2] DSM-5 Fact Sheets.
[3] Qwen2.5 tokenizer.

	Positive (Depressive)	Negative (Non-depressive)
NO. of Users	116	117
NO. of Texts	20,360	23,818
NO. of Tokens	1,024,978	1,115,951

Table 1: Dataset statistics.

events, and everyday stressors. Representative spans include *divorce*, *parents*, *school*, *teacher*, and *dropping out*.

- **Dimension 6: Language Use Patterns Related to Depression (secondary).** This dimension captures linguistic patterns such as negation, questions, and derogatory expressions. Representative spans include *I don't know*, *I don't like myself*, *I don't want to*, *I can't do it*, *why*, and *what should I do ?*.

Each user has multiple posts. We treat each post as the primary unit for review. Annotation is performed on semantic spans within each post. Each span is assigned to one of the six dimensions. A post may contain multiple spans, and different spans may belong to different dimensions.

Annotation was conducted by four senior researchers with expertise in psychology and depression scales. Annotators were blinded to the original binary labels in SWDD to ensure independent judgment. They periodically cross-checked overlapping samples to maintain consistency. Because the task relied more on qualitative judgment than on fixed quantitative criteria, the annotation guideline was not set in stone from the outset. It was revised in rounds as experts reviewed newly labeled data and resolved disagreements through discussion, and earlier annotations were corrected when the updated standards called for it. Given this evolving process, conventional inter-annotator agreement scores are not directly applicable. Since our annotation follows a different scheme from SWDD and all labeling was performed independently of the original labels, some user-level labels in our dataset differ from those in the original dataset.

This process yields the *CNSD Gold* set (100 positive and 100 negative users) and the *CNSD Test* set (16 positive and 17 negative users). Figure 1 summarizes the construction process.

Each entry (Figure 2) contains a binary label and a dimension-specific analysis. For each dimension, the analysis provides an overall assessment. It also lists post indices with brief justifications.

The overall statistics are presented in Table 1, and the 10,306 dimension-level annotations are summarized in Table 2.

Data Sample

User id: **user 61**

User posts:

post1: 'You are really planning to leave, I just don't believe it. Covering my eyes and closing my ears, pretending you will come back. Deceiving myself.',

post2, post3. post4, post5, ...

Label: **Depressed**

Dim 1: Depressive psychological state (primary)

Overall Assessment: The user presents a persistent low mood accompanied by a diminished sense of self-worth.

Post Selections: post_36 [self-negation], post_37 [persistent low mood], post_62 [self-negation], post_152 [nocturnal depressive mood], post_230 [self-harm ideation], post_278 [anxiety attack], post_275 [suicidal request], ...

Dim 2: Medical expressions related to depression (primary)

Overall Assessment: The posts mention medical help-seeking behaviors related to depression, such as clinical visits and antidepressant medication use.

Post Selections: post_318 [hypnotic medication use], post_321 [medication adherence], post_334 [distress during clinical follow-up], ...

Dim 3: Clinical symptoms related to depression (primary)

Overall Assessment: The posts mention depression-related clinical symptoms such as insomnia and fatigue.

Post Selections: post_173 [insomnia], post_209 [hypersomnia], post_273 [sleep disturbance], ...

Dim 4: Negative emotions (secondary)

Overall Assessment: The posts frequently express emotions such as sadness, hopelessness, and anger.

Post Selections: post_1 [extreme sadness], post_17 [hopelessness and helplessness], post_67 [crying], ...

Dim 5: Potential external causes of depression (secondary)

Overall Assessment: The posts mention external stressors, such as stressful life events.

Post Selections: post_94 [workplace criticism], post_101 [academic/school pressure], post_102 [social pressure], post_112 [social aversion], ...

Dim 6: Language Use Patterns Related to Depression (secondary)

Overall Assessment: The posts include crisis-oriented statements, such as "I don't know what to do" and "I want to die."

Post Selections: post_18 [suicidal ideation], post_203 [death wish], ...

Figure 2: A user-level annotated entry from CNSD-Gold (English translations).

Dimension	Negative User	Positive User
Dim1	117	2127
Dim2	2	555
Dim3	126	768
Dim4	514	2933
Dim5	193	863
Dim6	231	1877
Total	1183	9123

Table 2: Dimension-level annotation statistics for negative and positive users. This table reports the number of expert-annotated depression-related spans in each dimension. A single text may contain multiple spans.

4. Automated Dataset Generation Pipeline

Expert annotation requires substantial human effort and professional psychological expertise. As a result, user-level depression-risk datasets for Chinese social media remain scarce. To narrow this gap, we develop an automated dataset generation pipeline. We use it to construct *CNSD Silver* from the SWDD binary classification dataset. The pipeline is informed by the manually annotated *CNSD Gold* dataset (Section 3).

To build *CNSD Silver*, we randomly sample 100 positive (depression-risk) users and 100 negative (non-risk) users from SWDD. Each user contributes dozens to hundreds of posts. We keep users with at least 3,000 tokens in total.

The pipeline consists of two modules.

4.1. Module I: Dimension-Wise Automatic Labeling

Module I uses a mid-sized model (Qwen2.5-14B (Bai et al., 2023)) to automatically assign posts to the six depression-related dimensions ($\mathcal{D} = \{D_1, D_2, \ldots, D_6\}$). We fine-tune the model on text-level training data curated from positive users in *CNSD Gold*. After fine-tuning, the model can assign each post to one or more dimensions.

Data preparation for fine-tuning Module I.

1. For each dimension D_k, we extract posts from positive users in *CNSD Gold* that contain at least one annotated span of D_k. These posts form a set S_k.

2. We identify the dimension with the fewest training posts. The minimum count is $n_{\min} = 442$, which corresponds to *depression-related medical expressions*:

$$n_{\min} = \min_{k \in \{1,2,\ldots,6\}} n_k = 442.$$

3. To balance dimensions, we downsample each S_k to $n_{\min}$ when $n_k > n_{\min}$. This yields balanced sets S'_k:

$$S'_k = \begin{cases} S_k, & \text{if } n_k = n_{\min}, \\ \text{RandomSubset}(S_k, n_{\min}), & \text{if } n_k > n_{\min}. \end{cases}$$

We obtain 6×442 labeled positive posts. Each post is formatted using an instruction template such as: "This post belongs to [*category*] because it mentions [*evidence*]." We further augment the labels with 20 paraphrased expressions with the same meaning (Appendix A.2).

4. For negative training data, we split posts from negative users in *CNSD Gold* into individual texts. We remove any text that contains depression-related content in any dimension. From the remaining texts, we uniformly sample 6×442 examples. We assign them negative label expressions indicating that there is no evidence for the target dimension (As with the positive examples, we create 20 negative label expressions to reduce overfitting; details are provided in A.3).

5. We fine-tune Qwen2.5-14B with LoRA (Hu et al., 2022) on the combined set of $2 \times (6 \times 442)$ texts for 1 epoch. We use a learning rate of 5×10^{-5}, with the prompt provided in Appendix A.4.

4.2. Module II: Automatic Verification and Summarization

In Module II, we apply the fine-tuned Module I model to label all posts for each user. We then collect posts that are assigned to at least one dimension. Next, we use DeepSeek-R1-671B (DeepSeek-AI et al., 2025) to verify the assigned labels and the supporting evidence on a case-by-case basis. After verification, we summarize the collected posts using a few-shot prompting setup. The prompt instructs the model to act as a psychology expert and to produce structured summaries aligned with *CNSD Gold*. We also provide detailed requirements and gold-standard examples. The full prompt is shown in Appendix A.7.

5. Experiments

We design our experiments to answer two research questions. **First**, how effectively does our data generation pipeline produce high-quality structured annotations ? **Second**, how well do various large language models (LLMs) perform on user-level depression risk summarization and structured psychological analysis ?

Based on these questions, we conduct three sets of experiments:

1. To assess the quality of model-generated six-dimensional structured depression analyses. We compare models fine-tuned on *CNSD Gold* and *CNSD Silver* with baseline models and few-shot prompting.
2. To evaluate the performance of these fine-tuned models on user-level depression risk classification.
3. To benchmark leading LLMs on a structured depression analysis summarization task using *CNSD Gold*.

In Section 5.5, we also use *CNSD Gold* as a test set for binary depression classification. This further illustrates the dataset's versatility.

5.1. Baseline Models

We use the following generative baselines: DeepSeek-R1-Distill-14B (DeepSeek-AI et al., 2025), Qwen2.5-14B (QwenTeam, 2024), GPT-4o[4], GPT-4o-mini[5], DeepSeek-R1-671B[6] (DeepSeek-AI et al., 2025), and Llama3-8B-Chinese-Chat (Wang et al., 2024b). We summarize key characteristics in Table A.1.

5.2. Experimental Setup

We use NVIDIA A800 (80GB) and A100 (80GB) GPUs. Unless otherwise specified, we set the generation temperature to 0.7. All fine-tuning experiments use the LLaMA-Factory framework (Zheng et al., 2024). We apply LoRA (Hu et al., 2022) for 1 epoch with a learning rate of 5×10^{-5}.

Our experiments have two parts. The first evaluates the pipeline for data generation. The second benchmarks models on the *CNSD Gold* dataset for structured summarization and analysis generation.

For classification, we report accuracy, precision, recall, and F1-score. For generation, we report BLEU (Papineni et al., 2002), ROUGE-1 (Lin, 2004), and BERTScore (Zhang et al., 2020).

5.3. Task I: Data Generation

We conduct experiments on the *CNSD Test* dataset. We focus on two Qwen2.5-14B models fine-tuned on *CNSD Gold* and *CNSD Silver*:

- **Qwen2.5-14B Gold:** Fine-tuned on *CNSD Gold*.
- **Qwen2.5-14B Silver:** Fine-tuned on *CNSD Silver*. *CNSD Silver* is generated using the pipeline in Section 4 on 100 positive and 100 negative users sampled from SWDD.

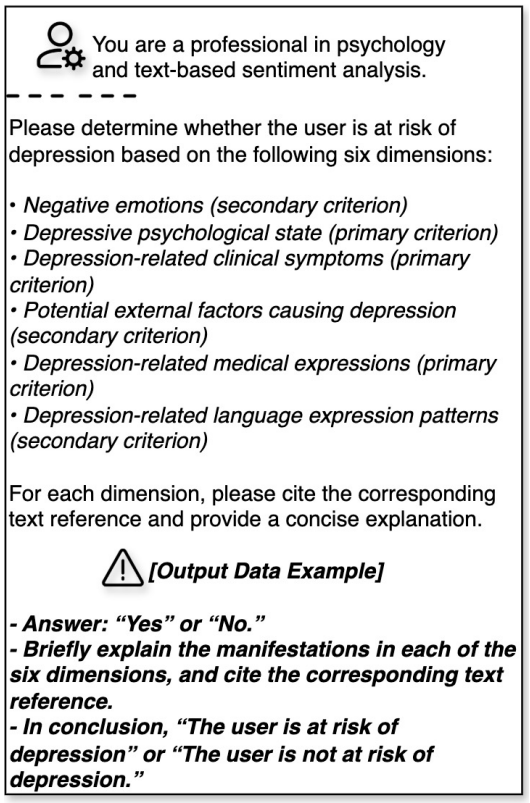


Figure 3: Prompt used for Table 3.

We compare these models with other baselines in two aspects:

1. Depression risk classification performance.
2. Quality of the generated six-dimensional analyses, including hallucinated or unsupported content.

Unless noted otherwise, we use user-level generation. We concatenate all posts from a user as input. We instruct the model to output a binary label and a corresponding justification.

5.3.1. Six-Dimensional Structured Depression Analysis

Prior work suggests that incorporating structured knowledge can improve LLM outputs (Moiseev et al., 2022). *CNSD Gold* and *CNSD Silver* provide fine-grained structured annotations across six dimensions. They support user-level summarization and analysis of social media content. We hypothesize that fine-tuning on these datasets can reduce hallucinations. It may also improve generation quality compared to few-shot prompting and larger models.

Table 3 compares different generation strategies using automatic metrics. Our Pipeline achieves

[4] https://openai.com/index/hello-GPT-4o/
[5] https://openai.com/index/GPT-4o-mini-advancing-cost-efficient-intelligence/
[6] https://www.deepseek.com/

Strategy	BERTScore	ROUGE-1	BLEU
Pipeline	**0.791**	**0.478**	**0.288**
FS:Qwen2.5 14B	0.649	0.076	0.075
FS:GPT-4o	0.678	0.269	0.094
FS:GPT-4o Mini	0.674	0.170	0.070
FS:DeepSeek R1 671B	0.678	0.301	0.054
FS:DeepSeek R1 Distill -14B	0.6756	0.191	0.073

Table 3: Comparison of generation strategies on text quality. In the table, FS stands for Few-Shot. Pipeline refers to the automated dataset generation pipeline proposed in Section 4, applied to our proposed dataset in Section 3. The purpose is to show that, with the same prompt, our pipeline achieves the highest text generation quality while meeting our task requirements. The prompt used in this experiment is shown in Figure 3.

the best scores on all three metrics: BERTScore (0.791), ROUGE-1 (0.478), and BLEU (0.288). These results indicate stronger semantic alignment, higher lexical overlap, and better n-gram precision. Few-shot (FS) prompting yields lower scores across models, including Qwen2.5-14B, GPT-4o, and DeepSeek-R1-671B. Notably, DeepSeek-R1-671B does not outperform the Pipeline. This suggests that the proposed Pipeline produces text that matches the reference more closely than few-shot prompting, across both smaller and larger models.

We also compare the original Qwen2.5-14B model with Qwen2.5-14B Gold and Qwen2.5-14B Silver (Table 4). We use two complementary evaluation methods.

- **Automatic metrics.** The original model achieves a BERTScore of 0.649. It increases to 0.764 for Gold (+16.9%) and 0.787 for Silver (+21.5%). ROUGE-1 rises from 0.076 to 0.217 for Gold (+172%) and 0.218 for Silver (+172%). BLEU improves from 0.075 to 0.186 for Gold (+149%) and 0.237 for Silver (+218%).

- **Human evaluation.** Two linguistic experts evaluate accuracy, coverage, and hallucination. Compared to the original Qwen2.5-14B, both Gold and Silver improve substantially. Silver increases by 33.4% in accuracy and 37.1% in coverage. Its hallucination score increases by 15.5% (higher is better). Gold achieves larger gains. Accuracy increases by 42.9% and coverage increases by 58.0%. The hallucination score increases by 26.8%. Overall, Gold performs best in human evaluation. Silver is slightly behind but remains close to Gold and clearly outperforms the original model.

In summary, Gold performs best in human evaluation. Silver performs better on automatic metrics. Both models substantially outperform the original model. These results show that fine-tuning with our dataset and pipeline can improve generation quality at semantic, lexical, and structural levels.

5.3.2. Classification Task

In the depression classification task, *Qwen2.5-14B Silver* achieves the best performance among all models. It reaches an accuracy of 0.944 and an F1 score of 0.941. It also surpasses *Qwen2.5-14B Gold*, which is fine-tuned on human-annotated data. Moreover, it outperforms large-scale models such as GPT-4o (accuracy = 0.917, F1 = 0.923) and DeepSeek-R1-671B (accuracy = 0.861, F1 = 0.872). These results support the effectiveness of our generation pipeline and suggest that the resulting silver data are of high quality. Detailed results are reported in Table 5.

5.4. Task II: Structured Summarization and Analysis Generation

Our dataset supports research on generative models for depression risk assessment from social media. It is particularly useful for user-level summarization of depressive signals. To the best of our knowledge, there is no directly comparable open-source dataset with the same annotation schema. We therefore conduct an initial benchmark of mainstream generative models on our annotated dataset of 233 users. Results are reported in Table 6.

We use the following prompt:

- **Instruction:** *Read the following texts written by one user. Then answer the question.*

- **Question:** Based on these texts, does this user exhibit a depressive mood ?

  - If the user exhibits a depressive mood, answer `Yes`.
  - If the user does not exhibit a depressive mood, answer `No`.

- **Output requirements:** Provide a brief justification based on evidence from the texts. Then produce a structured summary aligned with the annotation schema.

Model	BERTScore	ROUGE-1	BLEU	Human.acc	Human.cov	Human.hallu
Qwen2.5-14B	0.649	0.076	0.075	0.583	0.574	0.657
Qwen2.5-14B Silver	**0.787**	**0.218**	**0.237**	0.778	0.787	0.759
Qwen2.5-14B Gold	0.764	0.217	0.186	**0.833**	**0.907**	**0.833**

Table 4: Comparison of generation quality across models. `Qwen2.5-14B Gold` is fine-tuned on *CNSD Gold*. `Qwen2.5-14B Silver` is fine-tuned on *CNSD Silver*, which is generated by our pipeline (Section 4). We report automatic metrics (BERTScore, ROUGE-1, BLEU) and human evaluation metrics: *acc* (accuracy), *cov* (content coverage), and *hallu* (hallucination; higher is better). Higher values indicate better performance for all metrics.

Model	**Accuracy**	**Recall**	**Precision**	**F1**
Qwen2.5-14B	0.889	0.944	0.850	0.894
Qwen2.5-14B (FT, Silver)	**0.944**	0.889	**1.000**	**0.941**
Qwen2.5-14B (FT, Gold)	0.861	0.889	0.842	0.864
GPT-4o	0.917	**1.000**	0.857	0.923
GPT-4o-mini	0.667	**1.000**	0.600	0.750
DeepSeek-R1-671B	0.861	0.944	0.801	0.872
DeepSeek-R1-Distill-14B	0.889	**1.000**	0.818	0.900

Table 5: Comparison of Models for Generated Dataset Quality: Application to Classification Tasks.

The results vary across models. BERTScore falls in a narrow range (0.654–0.710), with GPT-4o achieving the highest score. ROUGE-1 is highest for DeepSeek-R1-671B (0.454), which exceeds the second-best model (GPT-4o at 0.346) by an absolute margin of 0.108. BLEU ranges from 0.093 (Qwen2.5-7B) to 0.152 (GPT-4o).

Model	BERT Score	ROUGE-1	BLEU
Llama3-8B	0.654	0.194	0.122
GLM4-9B-Chat	0.694	0.208	0.140
Qwen2.5-7B	0.680	0.208	0.093
DeepSeek-R1-671B	0.698	**0.454**	0.125
GPT-4o-mini	0.696	0.166	0.126
GPT-4o	**0.710**	0.346	**0.152**

Table 6: Model performance on the depression risk summarization and analysis generation task.

5.5. Task III: Classification Experiments

Our dataset is primarily designed for generation tasks. We also conduct a binary classification experiment for depression risk. In addition to our dataset, we use the original SWDD and WU3D datasets. For each dataset, we randomly sample 200 positive and 200 negative users.

Besides the generative models described in Appendix A.1, we evaluate BERT-based classifiers (Devlin et al., 2019). Implementation details are provided in Appendix A.5. Results are reported in Appendix A.6.

6. Conclusion

We introduce **CNSocialDepress**, a new dataset for depression risk analysis from Chinese social media. To our knowledge, it is among the first to provide both user-level and text-level labels. The dataset combines binary depression indicators with fine-grained six-dimensional analyses annotated by mental health professionals. We also develop an automated data generation pipeline. It reduces manual annotation effort and supports language model fine-tuning for classification and summarization. We expect **CNSocialDepress** to be useful for a range of research and applications, supporting early risk identification and downstream mental health interventions.

7. Limitations

While **CNSocialDepress** provides a valuable resource for depression risk analysis on Chinese social media, several limitations should be noted. First, the dataset is sourced exclusively from Weibo. This may introduce platform-specific demographic biases, such as the underrepresentation of rural or elderly populations.

Second, linguistic diversity is not fully captured. This includes dialectal variation and metaphorical expressions that are common in Chinese depressive discourse. This limitation may reduce generalizability across regions. Third, expert annotation ensures high quality, but it limits scalability. It also results in a relatively small dataset.

In addition, the current structured analyses focus primarily on depressive symptoms. They omit co-

morbid mental health conditions (e.g., anxiety) and broader social context. Future work will expand data collection to multiple platforms. We also plan to update lexicons to capture emerging expressions. Finally, we will explore hybrid annotation frameworks (e.g., expert-guided crowdsourcing) to improve coverage and efficiency. Privacy-preserving techniques and longitudinal tracking can further strengthen ethical and practical utility.

8. Ethics Statement

This dataset is built upon the SWDD benchmark (Cai et al., 2023), released in early 2023. The original data consist of user-level posts from Weibo. They were annotated by psychology experts based on DSM-5 criteria.

During re-annotation, we applied a second round of de-identification to protect privacy. We remove personal information, geographic locations, and other potential identifiers from the texts. We follow recommended data protection practices Benton et al. (2017) and comply with the GDPR (General Data Protection Regulation).[7] We also recruit experienced researchers in depression-related psychology to provide expert annotations.

This dataset is intended for research on depression risk identification and for building assistive models. It can support multi-dimensional analysis of risk-related signals and the generation of structured summaries. It is not designed for clinical diagnosis, triage, or treatment decisions. Any outputs should be used only as references for professionals or as a self-assessment aid.

No algorithmic system can replace in-person psychiatric evaluation or provide a definitive diagnosis. If a system built on this dataset flags a high-risk case, users should be encouraged to seek professional help. The same applies to individuals with persistent mood disturbances or impaired functioning. Online self-assessment tools cannot substitute for clinical evaluation.

9. Bibliographical References

Maryam Mohammed Aldarwish and Hafiz Farooq Ahmad. 2017. Predicting depression levels using social media posts. In *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*, pages 277–280. IEEE.

Shumaila Aleem, Noor ul Huda, Rashid Amin, Samina Khalid, Sultan S Alshamrani, and Abdullah Alshehri. 2022. Machine learning algorithms for depression: diagnosis, insights, and research directions. *Electronics*, 11(7):1111.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024. Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3250–3260, Torino, Italia. ELRA and ICCL.

Salma Almouzini, Asem Alageel, et al. 2019. Detecting arabic depressed users from twitter data. *Procedia Computer Science*, 163:257–265.

Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya, and Mueen Uddin. 2022. Deep learning for depression detection from textual data. *Electronics*, 11(5):676.

Priyanka Arora and Parul Arora. 2019. Mining twitter data for depression detection. In *2019 international conference on signal processing and communication (ICSC)*, pages 186–189. IEEE.

American Psychiatric Association. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21(21):591–643.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Krishna C Bathina, Marijn Ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. 2021. Individuals with depression express more distorted thinking on social media. *Nature human behaviour*, 5(4):458–466.

Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. 1996. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 67(3):588–597.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

André Bittar, Sumithra Velupillai, Angus Roberts, and Rina Dutta. 2019. Text classification to inform suicide risk assessment in electronic health records. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 40–44. IOS Press.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

[7] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504

Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480.

Ana-Maria Bucur, Andreea Moldovan, Krutika Parvatikar, Marcos Zampieri, Ashiqur Khudabukhsh, and Liviu Dinu. 2025. Datasets for depression modeling in social media: An overview. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 116–126, Albuquerque, New Mexico. Association for Computational Linguistics.

Fidel Cacheda, Diego Fernandez, Francisco J Novoa, and Victor Carneiro. 2019. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research*, 21(6):e12554.

Chengkun Cai, Haoliang Liu, Xu Zhao, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, John Lee, Jenq-Neng Hwang, and Lei Li. 2025a. Bayesian optimization for controlled image editing via llms. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chengkun Cai, Xu Zhao, Haoliang Liu, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, Jenq-Neng Hwang, and Lei Li. 2025b. The role of deductive and inductive reasoning in large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. 2023. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, 217:119538.

Junyeop Cha, Seoyun Kim, and Eunil Park. 2022. A lexicon-based approach to examine depression detection in social media: the case of twitter and university community. *Humanities and Social Sciences Communications*, 9(1):1–10.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Corinna Cortes. 1995. Support-vector networks. *Machine Learning*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephan Dreiseitl and Lucila Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359.

Zhiyuan Du, Yuxian Qian, Xiao Liu, Ming Ding, Yuxian Qian, Xiaoyang Liu, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335. Association for Computational Linguistics.

Egle Eensoo and Mathieu Valette. 2012. Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. In *TALN 2012*, volume 2, pages 367–374. GETALP-LIG.

SAGS Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Corina Benjet, Ronny Bruffaerts, Wai-Tat Chiu, Silvia Florescu, Giovanni de Girolamo, Oye Gureje, et al. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological medicine*, 48(9):1560–1571.

Andrea C Fernandes, Rina Dutta, Sumithra Velupillai, Jyoti Sanyal, Robert Stewart, and David Chandran. 2018. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, 8(1):7426.

Shang Gao, Mohammed Alawad, M Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B Durbin, Jennifer Doherty, Antoinette Stroup, et al. 2021. Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, 25(9):3596–3607.

GeminiTeam, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

GLMTeam, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Ruocheng Gu, Sen Jia, Yule Ma, Jinqin Zhong, Jenq-Neng Hwang, and Lei Li. 2025. Mocount: Motion-based repetitive action counting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9026–9034.

Yunchuan Guan, Yu Liu, Ke Zhou, Zhiqi Shen, Serge Belongie, Jenq-Neng Hwang, and Lei Li. 2025. Learning to learn weight generation via local consistency diffusion. *arXiv preprint arXiv:2502.01117*. Accepted to CVPR 2026.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Zhihua Guo, Nengneng Ding, Minyu Zhai, Zhenwen Zhang, and Zepeng Li. 2023. Leveraging domain knowledge to improve depression detection on chinese social media. *IEEE Transactions on Computational Social Systems*, 10(4):1528–1536.

Mika Hämäläinen, Pattama Patpong, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Detecting depression in Thai blog posts: a dataset and a baseline. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 20–25, Online. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Khan Md Hasib, Md Rafiqul Islam, Shadman Sakib, Md Ali Akbar, Imran Razzak, and Mohammad Shafiul Alam. 2023. Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey. *IEEE Transactions on Computational Social Systems*, 10(4):1568–1586.

Qi He and Chunyu Qu. 2025a. Modular landfill remediation for ai grid resilience. *arXiv preprint*.

Qi He and Chunyu Qu. 2025b. Waste-to-energy-coupled ai data centers: Cooling efficiency and grid resilience. *arXiv preprint*.

Qi He, Rui Shan, Chunyu Qu, Yuan Tang, and Yue Zou. 2026. Rare-earth exposure and bottlenecks in ai data centers: Cost and schedule risk. *SSRN Electronic Journal*. Available at SSRN.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12.

Shaoxiong Ji, Tianyu Zhang, Laith Ansari, Jie Fu, Piyush Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190. European Language Resources Association.

Sen Jia and Lei Li. 2024. Adaptive masking enhances visual grounding. *arXiv preprint arXiv:2410.03161*.

Zhongyu Jiang, Wenhao Chai, Lei Li, Zhuoran Zhou, Cheng-Yen Yang, and Jenq-Neng Hwang. 2025. Unihpr: Unified human pose representation via singular value contrastive learning. *arXiv preprint arXiv:2510.19078*.

Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang. 2024. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Heysem Kaya, Dmitrii Fedotov, Denis Dresvyanskiy, Metehan Doyran, Danila Mamontov, Maxim Markitantov, Alkim Almila Akdag Salah, Evrim Kavcar, Alexey Karpov, and Albert Ali Salah. 2019. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 27–35.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, and Yong Li. 2024. Depression detection on social media with large language models. *arXiv preprint arXiv:2403.10750*.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

Genghao Li, Bing Li, Langlin Huang, Sibing Hou, et al. 2020. Automatic construction of a depression-domain lexicon based on microblogs: text mining study. *JMIR medical informatics*, 8(6):e17650.

Hao Li, Ju Dai, Xin Zhao, Feng Zhou, Junjun Pan, and Lei Li. 2025a. Wav2sem: Plug-and-play audio semantic decoupling for 3d speech-driven facial animation. *arXiv preprint arXiv:2505.23290*.

Hao Li, Ju Dai, Feng Zhou, Kaida Ning, Lei Li, and Junjun Pan. 2025b. Au-blendshape for fine-grained stylized 3d facial expression manipulation. *arXiv preprint arXiv:2507.12001*.

Lei Li, Sen Jia, and Jenq-Neng Hwang. 2026. Multiple human motion understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6297–6305.

Lei Li, Sen Jia, Jianhao Wang, Zhaochong An, Jiaang Li, Jenq-Neng Hwang, and Serge Belongie. 2025c. Chatmotion: A multimodal multi-agent for human motion analysis. *arXiv preprint arXiv:2502.18180*.

Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, and Jenq-Neng Hwang. 2025d. Human Motion Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lei Li, Tianfang Zhang, Zhongfeng Kang, and Xikun Jiang. 2022a. Mask-fpan: Semi-supervised face parsing in the wild with de-occlusion and uv gan. *arXiv preprint arXiv:2212.09098*.

Mingzheng Li, Haojie Xu, Weifeng Liu, and Jiangwei Liu. 2022b. Bidirectional lstm and attention for depression detection on clinical interview transcripts. In *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, pages 638–643. IEEE.

Zepeng Li, Zhengyi An, Wenchuan Cheng, Jiawei Zhou, Fang Zheng, and Bin Hu. 2023. Mha: a multimodal hierarchical attention model for depression detection in social media. *Health information science and systems*, 11(1):6.

Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 407–411.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Libin Liu, Shen Chen, Sen Jia, Jingzhe Shi, Zhongyu Jiang, Can Jin, Wu Zongkai, Jenq-Neng Hwang, and Lei Li. 2024. Graph canvas for controllable 3d scene generation. *arXiv preprint arXiv:2412.00091*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ashwag Maghraby and Hosnia Ali. 2022. Modern standard arabic mood changing and depression dataset. *Data in Brief*, 41:107999.

Anshu Malhotra and Rajni Jindal. 2022. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, 130:109713.

Keshu Malviya, Bholanath Roy, and SK Saritha. 2021. A transformers approach to detect depression in social media. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 718–723. IEEE.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured knowledge infusion for large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, Seattle, United States. Association for Computational Linguistics.

Josiane Mothe, Faneva Ramiandrisoa, and Md Zia Ullah. 2022. Comparison of machine learning models for early depression detection from users' posts. In *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*, volume 1018 of *Studies in Computational Intelligence book series (SCI)*, pages 111–139. Springer International Publishing.

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, Mike Conway, et al. 2017. Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study. *Journal of medical Internet research*, 19(2):e6895.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 88–97.

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.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2022. Overview of erisk at clef 2022: Early risk prediction on the internet (extended overview). In *CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy*, Bologna, Italy. CEUR Workshop Proceedings. CC BY 4.0; ISSN: 1613-0073.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bénédicte Pincemin. 2022. Sémantique textométrique. *La sémantique au pluriel. Théories et méthodes*, pages 373–396.

QwenTeam. 2024. Qwen2.5: A party of foundation models.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. MentalHelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203, Torino, Italia. ELRA and ICCL.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Alba M Mármol Romero, Adrián Moreno Muñoz, Flor Miriam Plaza Del Arco, M Dolores Molina-González, María Teresa Martín Valdivia, L Alfonso Urena Lopez, and Arturo Montejo Ráez. 2024. Mentalriskes: A new corpus for early detection of mental disorders in spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11204–11214.

Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

André Salem. 1986a. Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure*, 1:5–28.

André Salem. 1986b. Segments répétés et analyse statistique des données textuelles. *Histoire & mesure*, pages 5–28.

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2024. Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, 58(1):273–300.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Elizabeth M Seabrook, Margaret L Kern, Ben D Fulcher, and Nikki S Rickard. 2018. Predicting depression from language-based emotion dynamics: longitudinal analysis of facebook and twitter status updates. *Journal of medical Internet research*, 20(5):e168.

Faisal Muhammad Shah, Farzad Ahmed, Sajib Kumar Saha Joy, Sifat Ahmed, Samir Sadek, Rimon Shil, and Md Hasanul Kabir. 2020. Early depression detection from social network using deep learning techniques. In *2020 IEEE region 10 symposium (TENSYMP)*, pages 823–826. IEEE.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat Seng Chua, and Wendy Hall. 2018. Cross-domain depression detection via harvesting social media. In *Proceedings of the International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Jingzhe Shi, Qinwei Ma, Hongyi Liu, Hang Zhao, Jeng-Neng Hwang, and Lei Li. 2025. Explaining context length scaling and bounds for language models. *arXiv preprint arXiv:2502.01481*.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. 2024. Scaling law for time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pengcheng Shi, Jiawei Chen, Jiaqi Liu, Xinglin Zhang, Tao Chen, and Lei Li. Medal s: Spatio-textual prompt model for medical segmentation. In *CVPR 2025: Foundation Models for 3D Biomedical Image Segmentation*.

Sajad Sotudeh, Hanieh Deilamsalehy, Franck Dernoncourt, and Nazli Goharian. 2021. TLDR9+: A large scale resource for extreme summarization of social media posts. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 142–151, Online and in Dominican Republic. Association for Computational Linguistics.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. MentSum: A resource for exploring summarization of mental health online posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2682–2692, Marseille, France. European Language Resources Association.

Matthew Squires, Xiaohui Tao, Soman Elangovan, Raj Gururajan, Xujuan Zhou, U Rajendra Acharya, and Yuefeng Li. 2023. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics*, 10(1):10.

Maxim Stankevich, Ivan Smirnov, Natalia Kiselnikova, and Anastasia Ushakova. 2020. Depression detection from social media profiles. In *Data Analytics and Management in Data Intensive Domains: 21st International Conference, DAMDID/RCDL 2019, Kazan, Russia, October 15–18, 2019, Revised Selected Papers 21*, pages 181–194. Springer.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and

computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

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*.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018a. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018b. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF (working notes)*.

Qamar Un Nisa and Rafi Muhammad. 2021. Towards transfer learning using bert for early detection of self-harm of social media users. *Proceedings of the Working Notes of CLEF*, pages 21–4.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jifei Wang, Zhenping Zhao, Jing Yang, Limin Wang, Mei Zhang, and Maigeng Zhou. 2024a. The association between depression and all-cause, cause-specific mortality in the chinese population—china, 2010–2022. *China CDC Weekly*, 6(40):1022.

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. 2024b. Llama3-8b-chinese-chat (revision 6622a23).

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. 2020a. A multimodal feature fusion-based method for individual depression detection on sina weibo. In *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*, pages 1–8.

Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. 2020b. A multimodal feature fusion-based method for individual depression detection on sina weibo. In *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*, pages 1–8.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024c. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp. In *Proceedings of the first international workshop on language cognition and computational models*, pages 11–21.

Min Yen Wu, Chih-Ya Shen, En Tzu Wang, and Arbee LP Chen. 2020. A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, 54(2):225–244.

Dong Xue* Xin Yan. 2023. Mindchat: Psychological large language model. `https://github.com/X-D-Lab/MindChat`.

Jinyuan Xu, Tian Lan, Mathieu Valette, Pierre Magistry, and Lei Li. 2025. TinyMentalLLMs enable depression detection in Chinese social media texts. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1352–1363, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li,

Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Juhua Yang and Lulu Li. 2009. Intergenerational dynamics and family solidarity: A comparative study of mainland china, japan, korea and taiwan. *Sociological Studies*, 3:26–53.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Tingting Yang, Fei Li, Donghong Ji, Xiaohui Liang, Tian Xie, Shuwan Tian, Bobo Li, and Peitong Liang. 2021. Fine-grained depression analysis based on chinese micro-blog reviews. *Information Processing & Management*, 58(6):102681.

Wenhao Yang, Jianguo Wei, Wenhuan Lu, and Lei Li. 2025. You only speak once to see. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Xingwei Yang, Rhonda McEwen, Liza Robee Ong, and Morteza Zihayat. 2020. A big data analytics framework for detecting user-level depression from social networks. *International Journal of Information Management*, 54:102141.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Ziyu Yao, Xuxin Cheng, Zhiqi Huang, and Lei Li. 2025. CountLLM: Towards Generalizable Repetitive Action Counting via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Hossein Yazdavar, Mohammad Saeid Mahdavinejad, Goonmeet Bajaj, William Romine, Amit Sheth, Amir Hassan Monadjemi, Krishnaprasad Thirunarayan, John M Meddar, Annie Myers, Jyotishman Pathak, et al. 2020. Multimodal mental health analysis in social media. *Plos one*, 15(4):e0226248.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models.

Yutaka Miyaji Yuka Niimi. 2021. Machine learning approach for depression detection in Japanese. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 346–353, Shanghai, China. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Yang, and Guanghui Fu. 2024. Chinese MentalBERT: Domain-adaptive pre-training on social media for Chinese mental health text analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10574–10585, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Yipeng Zhang, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, and Jiebo Luo. 2021. Monitoring depression trends on twitter during the covid-19 pandemic: observational study. *JMIR infodemiology*, 1(1):e26769.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. Symptom identification for

interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Han-yu Zhou, Wen-qi Zhu, Wen-yi Xiao, Ya-ting Huang, Kang Ju, Hong Zheng, and Chao Yan. 2023. Feeling unloved is the most robust sign of adolescent depression linking to family communication patterns. *Journal of Research on Adolescence*, 33(2):418–430.

Zhuoran Zhou, Zhongyu Jiang, Wenhao Chai, Cheng-Yen Yang, Lei Li, and Jenq-Neng Hwang. 2024. Efficient domain adaptation via generative prior for 3d infant pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

A. Appendix

A.1. Baseline Models

- **DeepSeek-R1** is a large-scale model designed for complex reasoning. It performs well on tasks such as mathematics and programming, while remaining competitive on general-purpose benchmarks. We include it as a strong reasoning-oriented baseline for structured psychological analysis and depression risk assessment.
- **DeepSeek-R1-Distill-14B** is a distilled reasoning model derived from Qwen2.5-14B via knowledge distillation. It aims to retain key reasoning capabilities at lower computational cost. We include it as an efficient baseline for depression risk analysis.
- **Qwen2.5-14B** is a large language model from the Qwen series. It provides strong instruction following and long-form generation. We include it as a Chinese-capable baseline for generating structured psychological analyses from social media posts.
- **GPT-4o** is a multimodal large language model released by OpenAI. It supports multilingual generation, including Chinese. We include it as a strong general-purpose baseline for depression detection and psychological analysis.
- **GPT-4o-mini** is a smaller model in the GPT-4o family. It is designed for efficiency and lower inference cost. We include it as a lightweight baseline for depression risk identification and structured summarization.
- **Llama3-8B-Chinese-Chat**, derived from Meta-Llama-3-8B, is an instruction-tuned model for Chinese dialogue. It supports both Chinese and English inputs. We include it as a Chinese dialogue-oriented baseline for user-level depression risk analysis and generation.

A.2. 20 Randomly Generated Expression for the Positive Users

1. Carefully read the complete user Weibo text below and extract key information across six specified dimensions according to the content. Also pay attention to the user's depressive state, ensuring that each dimension's explanation is consistent with this state.
   [Input] User depressive state: {label}, Full Weibo text: {text}.
   [Output requirements] - For each dimension, if evidence exists in the text, provide the text index (e.g., text_xx, 文本_xx) and a brief note

(e.g., text_xx[sad]). - If no evidence is found for a dimension, state clearly that no relevant evidence was found.
Note: Only output text index and explanation, not the actual text.

2. Read the entire user Weibo text below and, based on its content, extract core information under the following six dimensions. Also consider the user's depressive status, ensuring that each dimension's description matches this state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - For each dimension, if relevant information is detected, provide the text index (e.g., text_xx, 文本_xx) and a short note (e.g., text_xx[unhappy]). - For dimensions without evidence, explicitly state no relevant evidence found.
Note: Output should include only text index and explanation, not the actual text.

3. Read through the complete user Weibo text below and extract key information across six dimensions based on its content. Pay close attention to the user's depressive state, ensuring that the explanations are consistent with it.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If relevant evidence is found in a dimension, provide the text index (e.g., text_xx, 文本_xx) and a short note (e.g., text_xx[sorrow]). - If no evidence is found, indicate explicitly.
Note: Provide only indices and explanations, not the original content.

4. Carefully read the full Weibo text below and extract key evidence across six dimensions according to the text. At the same time, take into account the user's depressive state, ensuring that explanations for each dimension align with this state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - For each dimension, if there is corresponding evidence, output the text index (e.g., text_xx, 文本_xx) and a related note (e.g., text_xx[frustrated]). - If no evidence is found, state that no evidence exists.
Note: Only provide text indices and notes, not the raw text.

5. Please read the full Weibo text below and extract key information from six dimensions. At the same time, pay attention to the user's depressive state, ensuring that your explanations match this state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - For each dimension, if evidence is detected, output the text index (e.g., text_xx, 文本_xx) and a short description (e.g., text_xx[sadness]). - If no evidence is found, specify that no evidence was found.
Note: Only output text indices and explanations, not the original text.

6. Read the complete Weibo text below carefully and extract key information across six dimensions based on the text. Pay attention to the user's depressive state and ensure that your descriptions are consistent with it.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If relevant evidence is found, provide the text index (e.g., text_xx, 文本_xx) and a brief description (e.g., text_xx[low mood]). - If no evidence is found, clearly state that no evidence exists.
Note: Only provide text indices and explanations, not the original text.

7. Please read through the complete user Weibo text below and extract key information across six dimensions based on its content. Pay attention to the user's depressive state, ensuring that each explanation is aligned with this state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - For each dimension, if evidence exists, provide the text index (e.g., text_xx, 文本_xx) and a short note (e.g., text_xx[loss]). - If no evidence is found, clearly state no relevant evidence.
Note: Only provide text indices and notes, not the original text.

8. Read the Weibo text below and extract key evidence across six dimensions according to the content, while considering the user's depressive state to ensure explanations are consistent.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If a dimension contains relevant evidence, provide the text index (e.g., text_xx, 文本_xx) and a short note (e.g., text_xx[low mood]). - If no evidence is found, state explicitly that no evidence exists.
Note: Output only text indices and explanations, not the original text.

9. Read the complete Weibo text carefully and extract key evidence across six dimensions based on the text. Pay attention to the user's depressive state, ensuring that explanations match this state.
[Input] User depressive state: {label}, Full

Weibo text: {text}.
[Output requirements] - For each dimension, if relevant evidence is found, output the text index (e.g., text_xx, 文本_xx) and a brief explanation (e.g., text_xx[depressed]). - If no evidence is found, state explicitly that there is none.
Note: Provide only text indices and notes, not the original text.

10. Read through the entire Weibo text and extract core information across six dimensions according to the content. Pay attention to the user's depressive state and ensure that the descriptions are consistent with this state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - For each dimension, if relevant evidence exists, output the text index (e.g., text_xx, 文本_xx) and a short explanation (e.g., text_xx[pessimistic]). - If no evidence is found, state clearly that no evidence exists.
Note: Only output text indices and explanations, not the original content.

11. Please read the user Weibo text below completely and extract essential information across six dimensions. Ensure that the explanations for each dimension align with the depressive state given.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If relevant evidence is detected, list the text index (e.g., text_xx, 文本_xx) and a short description (e.g., text_xx[melancholy]). - If no evidence is found, specify that no evidence was detected.
Note: Output should only contain indices and notes, not the actual text.

12. Read carefully the full Weibo text provided and extract key evidence from six dimensions. Pay attention to the user's depressive state, ensuring that outputs remain consistent with this state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If a dimension contains evidence, output the text index (e.g., text_xx, 文本_xx) and a concise note (e.g., text_xx[disheartened]). - If no evidence exists, specify so.
Note: Output indices and explanations only, not original content.

13. Please go through the complete Weibo text below and identify critical information across six dimensions. Ensure the notes correspond with the user's depressive condition.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - Where evidence is found, list the text index (e.g., text_xx, 文本_xx) with a short explanation (e.g., text_xx[downhearted]). - If no evidence is detected, indicate explicitly.
Note: Provide only text indices and notes, excluding original text.

14. Read carefully the following Weibo text and extract evidence across six preset dimensions. Make sure the explanation for each aligns with the depressive state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If evidence exists for a dimension, output the text index (e.g., text_xx, 文本_xx) and a short remark (e.g., text_xx[low mood]). - If no evidence is found, specify clearly.
Note: Output only indices and remarks, not the raw text.

15. Read through the following Weibo text and extract major evidence across six given dimensions. Ensure the output is consistent with the depressive state specified.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If evidence is found, provide the text index (e.g., text_xx, 文本_xx) with a brief note (e.g., text_xx[hopeless]). - If no evidence appears, state clearly.
Note: Only provide indices and notes, not original text.

16. Please carefully review the Weibo text below and extract key information from six preset dimensions. All notes must align with the depressive condition of the user.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If there is evidence, output the corresponding text index (e.g., text_xx, 文本_xx) and a short remark (e.g., text_xx[emotional slump]). - If not found, specify clearly that none exists.
Note: Only include indices and notes, no original content.

17. Please go through the complete Weibo text and extract the critical information across six given dimensions. Make sure your notes are consistent with the depressive state provided.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - For each dimension, if evidence is found, provide the text index (e.g., text_xx, 文本_xx) with a concise explanation

(e.g., text_xx[feeling low]). - If no evidence exists, indicate so.
Note: Only output indices and explanations, not actual text.

18. Carefully read the Weibo text provided and extract important evidence across six specified dimensions. Keep the explanations aligned with the depressive state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If a dimension contains evidence, state the text index (e.g., text_xx, 文本_xx) and a short remark (e.g., text_xx[upset]). - If not, state that no evidence is found.
Note: Output only indices and remarks, not text.

19. Read the Weibo text completely and extract relevant evidence from six dimensions. Ensure each dimension's note is consistent with the depressive state given.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If relevant evidence is detected, provide the text index (e.g., text_xx, 文本_xx) and short description (e.g., text_xx[distressed]). - If not found, state clearly no evidence.
Note: Only indices and notes, exclude text itself.

20. Please read through the following full Weibo text and extract essential evidence across six preset dimensions. Ensure all notes are consistent with the depressive state.
[Input] User depressive state: {label}, Full Weibo text: {text}.
[Output requirements] - If evidence is found, provide the text index (e.g., text_xx, 文本_xx) and a short explanation (e.g., text_xx[worry]). - If no evidence is detected, state that explicitly.
Note: Only indices and explanations, not original text.

A.3. 20 Distinct Label Expressions for the Negative Users

1. "This text does not contain any expressions of negative emotions, nor does it involve any dimensions of depression."
2. "There are no descriptions related to negative emotions in the text, so it does not fit any depression-related dimensions."
3. "This text does not include expressions of negative emotions, and therefore it is not classified under any depression dimensions."
4. "The content does not reflect negative emotions and does not belong to any depression indicators."
5. "This passage contains no expressions of negative emotions and is not applicable to any depression dimensions."
6. "There are no descriptions of negative emotions in the text, so it is not classified under any depression-related dimensions."
7. "This content does not exhibit negative emotions and therefore does not involve any depression dimensions."
8. "The text does not contain any expressions related to negative emotions, nor does it meet the criteria for depression dimensions."
9. "There are no signs of negative emotions in this text, so it does not belong to any depression dimension."
10. "The text does not describe negative emotions and does not cover any depression-related dimensions."
11. "This passage does not contain any expressions related to negative emotions and is not within the scope of depression dimensions."
12. "The content does not include expressions of negative emotions, so it does not fall into any depression dimensions."
13. "There are no expressions of negative emotions in the text, nor does it meet the criteria for depression dimensions."
14. "This text does not exhibit any negative emotions and does not involve any depression dimensions."
15. "This content does not include descriptions of negative emotions and therefore is not classified under any dimensions of depression."
16. "The text does not show any negative emotions and is not applicable to depression-related dimensions."
17. "No expressions of negative emotions are seen in this passage, so it does not meet any depression dimension standards."
18. "There are no expressions of negative emotions in the text, so it does not belong to any depression dimension."
19. "This text does not display any expressions related to negative emotions and does not cover the dimensions of depression."

20. "The content lacks descriptions of negative emotions and is therefore not classified under any depression dimensions."

A.4. Instructions for the Fine-Tuning Process

- **Complete Instruction = Task Instruction + Supplementary Instruction**

Task Instruction:
This task involves the following 6 dimensions of depression:

- Potential External Causes of Depression (Secondary Judgment Criterion)
- Depression-Related Clinical Symptoms (Primary Judgment Criterion)
- Depression-Related Language Expression Patterns (Secondary Judgment Criterion)
- Depression-Related Medical Expressions (Primary Judgment Criterion)
- Depressive Psychological State (Primary Judgment Criterion)
- Negative Emotions (Secondary Judgment Criterion)

Supplementary Instructions (randomly select one):

1. "Please read the following text segment and determine whether it contains any of the above 6 expressions of depressive emotion, and provide a brief explanation; if none is present, please indicate that the text does not belong to any depression dimension."
2. "Read the following text and analyze whether it demonstrates any one of the aforementioned 6 depressive emotion expression dimensions, and provide a brief explanation; if such expression is absent, please indicate that the text does not correspond to any dimension."
3. "Please carefully read the following text segment and confirm whether any one of the 6 depressive emotion expression dimensions mentioned above exists, and include a brief explanation; if not, please state that the text does not belong to any dimension."
4. "Please read the following content and determine whether it embodies any one of the above 6 depressive emotion expression dimensions, and provide a brief explanation; if not, please indicate that the text does not cover any depression dimension."
5. "Read the following text segment and determine whether it contains any one of the 6 depressive emotion expressions mentioned above, and provide a concise explanation; if not, please state that the text does not belong to any depression dimension."
6. "Please read the following text and determine whether any one of the above 6 depressive emotion expression dimensions appears, and include a brief explanation; if not, please indicate that the text does not meet any dimension."
7. "Read the following text and check whether any one of the above 6 depressive emotion expression dimensions is presented, and provide a brief explanation; if not, please state that the text does not belong to any dimension."
8. "Please review the following text segment and determine whether any one of the above 6 depressive emotion expression dimensions exists, and provide a brief explanation; if not, then indicate that the text does not belong to any dimension."
9. "Read the following text and confirm whether any one of the above 6 depressive emotion expression dimensions is embodied, and provide a brief explanation; if not, please state that the text does not involve any depression dimension."
10. "Please read the following content and determine whether it contains any one of the 6 depressive emotion expression dimensions mentioned earlier, and provide a brief explanation; if not, please indicate that the text does not belong to any dimension."
11. "Read the following text segment and verify whether it demonstrates any one of the above 6 depressive emotion expression dimensions, and include a brief explanation; if not, please state that the text does not meet any dimension."
12. "Please read the following text and check whether it exhibits any one of the above 6 depressive emotion expression dimensions, and provide a brief explanation; if not, please indicate that the text does not cover any depression dimension."
13. "Read the following text and confirm whether any one of the 6 depressive emotion expressions mentioned above exists, and include a brief explanation; if such a feature is absent, please state that the text does not belong to any dimension."

14. "Please carefully read the following text segment and determine whether it embodies any one of the above 6 depressive emotion expression dimensions, and provide a concise explanation; if not, please indicate that the text does not correspond to any dimension."

15. "Read the following text segment and confirm whether it contains any one of the above 6 depressive emotion expression dimensions, and provide a brief explanation; if not, please state that the text does not involve any depression dimension."

16. "Please read the following text segment and determine whether any one of the 6 depressive emotion expression dimensions mentioned above appears in the text, and include a brief explanation; if not, please indicate that the text does not belong to any dimension."

17. "Read the following content and check whether it contains any one of the above 6 depressive emotion expression dimensions, and provide a brief explanation; if not, please state that the text does not belong to any depression dimension."

18. "Please read the following content and determine whether any one of the above 6 depressive emotion expression dimensions exists, and provide a brief explanation; if not, please indicate that the text does not involve any dimension."

19. "Read the following text segment and confirm whether it demonstrates any one of the above 6 depressive emotion expression dimensions mentioned above, and provide a brief explanation; if not, please state that the text does not meet any dimension."

20. "Please read the following text and determine whether it contains any one of the above 6 depressive emotion expression dimensions, and provide a concise explanation; if not, please indicate that the text does not belong to any depression dimension."

A.5. BERT-based Models

BERT-base-chinese[8] follows the BERT-Base architecture (Devlin et al., 2019) and is pre-trained on large-scale Chinese corpora. It uses masked language modeling (MLM) and next sentence prediction (NSP). The model learns bidirectional representations for Chinese text, typically at the character level.

Chinese-RoBERTa-wwm-ext (Cui et al., 2019) is based on RoBERTa (Liu et al., 2019b) and uses whole word masking (WWM). It masks complete words rather than individual characters during MLM. This design can better capture word-level semantics in Chinese. The extended pre-training ("ext") on large Chinese corpora further improves robustness across NLP tasks.

StructBERT-mental (Wang et al., 2019) builds on StructBERT, which incorporates structural objectives during pre-training. It emphasizes word- and sentence-level structure. In addition, it is adapted for mental health text, which can benefit mental health-related classification.

[8] https://huggingface.co/google-bert/bert-base-chinese

A.6. Classification Performance Comparison

Model/Test Dataset	Swdd Origin		Wu3d		CNSD (Gold)	
	Acc	F1	Acc	F1	Acc	F1
RoBERTa Chinese (trained by swdd origin)	0.89	0.90	0.76	0.80	0.88	0.87
RoBERTa Chinese (trained by wu3d)	0.88	0.87	0.90	0.91	0.82	0.78
Bert based Chinese (trained by swdd origin)	0.89	0.90	0.76	0.80	0.85	0.82
Bert based Chinese (trained by wu3d)	0.88	0.87	0.90	0.91	0.83	0.80
StructBERT-mental (trained by swdd origin)	0.81	0.84	0.69	0.76	0.85	0.86
StructBERT-mental (trained by wu3d)	0.85	0.85	**0.92**	**0.92**	0.83	0.80
Llama3-8b	0.86	0.86	0.79	0.79	0.60	0.63
Glm4-9b-chat	0.76	0.81	0.77	0.81	0.79	0.82
Qwen2.5 7b	0.88	0.89	0.84	0.85	**0.94**	**0.94**
GPT-4o-Mini	0.65	0.74	0.65	0.74	0.53	0.68
GPT-4o	**0.92**	**0.92**	0.91	0.92	0.83	0.86
DeepSeek-R1 671b	0.86	0.87	0.89	0.90	0.81	0.84

Table 7: Model Classification Performance Comparison on Different Test Datasets.

A.7. Module II Prompt

You are a professional psychologist and text sentiment analysis expert. Below is a set of sentence-by-sentence analysis results generated by a smaller model concerning all of the user's Weibo posts, each result presented in the format "Text Label: Analysis Content." Please use your professional knowledge to review, filter, and correct these analysis contents, and then produce a final comprehensive analysis report.

[Task Requirements]

1. **Review and filter** the sentence-by-sentence analysis provided by the small model, selecting those text entries **that are useful** for generating the final report.

2. **Check** whether there is any classification error within these entries. If you discover that the small model has incorrectly assigned certain texts to an incorrect dimension (for example, **classifying "unhappy" incorrectly** under depression-related medical expressions), please **correct** it based on the actual semantics and psychological knowledge, reassign it to a more reasonable dimension, or exclude that erroneous information from the final report.

3. The user's depression status is **[already known depression label]**. Please ensure that the final comprehensive analysis report **is consistent** with this label:

    - If the label is **"depressed,"** each dimension in the report should **fully reflect** evidence related to depression;
    - If the label is **"non-depressed,"** then the report should **not contain** extensive mentions of depressive features.

4. In your descriptions, please **avoid using specific time quantifiers** (such as "≥2 weeks"), and instead use a neutral description such as **"multiple occurrences."**

5. The **final report** should be **organized under the following six dimensions**:

    - Negative emotions (secondary judgment criterion)
    - Depressive psychological state (primary judgment criterion)
    - Depressive clinical symptoms (primary judgment criterion)
    - Potential external causes of depression (secondary judgment criterion)
    - Depression-related medical expressions (primary judgment criterion)
    - Depression-related language expression patterns (secondary judgment criterion)

    In each dimension, please **cite the corresponding text label** and provide a concise explanation, ensuring that all classifications in the final report are **your corrected judgments**. The output **must not contain** any references to internal corrections (such as "already removed" or "already fixed") or references to **specific error rates**.

[Reference Gold Standard Answers Example]
[Positive Example]
{positive_example}

[Negative Example]
{negative_example}

[Sample Input Data]
User's known depression label: **[{label}]**
Small model's sentence-by-sentence analysis results:
{input}

[Task]
Please, based on the requirements above, **review, filter, and correct** the sentence-by-sentence analysis provided by the small model, and then produce a **final comprehensive analysis report**. The report must be **organized under six dimensions**, and in each dimension, please **cite the corresponding text label** along with a concise explanation. **Ensure** that the final report, after your corrections, **is consistent** with the user's known label **[[already known depression label]]** , and that it **directly reflects** the corrected chain of evidence and judgment conclusions, **without** including internal correction statements or specific time descriptors.

Now, please produce the final comprehensive analysis report.

Figure 4: Module II Prompt.

Depression detection in Modern Greek

Vivian Stamou[1], **George Mikros**[2], **George Markopoulos**[1], **Spyridoula Varlokosta**[1]

[1]National and Kapodistrian University of Athens, Greece,
[2]Hamad Bin Khalifa University, Qatar, Affiliation3
Panepistimiopoli, Zografou 157 72
Education City, Doha, Qatar 34110
vivianstamou@gmail.com, gmikros@hbku.edu.qa,
{gmarkop, svarlokosta}@phil.uoa.gr

Abstract

Despite advancements in NLP-based mental health screening, research remains predominantly English-centric, leaving under-resourced languages insufficiently explored. This study investigates depression detection in Modern Greek social media through a series of experiments. We benchmark traditional machine learning (ML) models against transformer architectures (GreekBERT, GreekSocialBERT, mBERT, and XLM-R) under two settings: a topic-oriented control corpus and a high-similarity stress-test contrasting a gold case of a depressed user with a matched control. Transformer models consistently outperform ML models (F1 = 0.95) but offer limited interpretability. To address this limitation, we incorporate LIWC-derived psycholinguistic features with SHAP explanations to examine model behavior in relation to established linguistic markers. The analysis reveals linguistic patterns consistent with depressive symptoms, such as reduced work-related engagement, social withdrawal, and the motivational deficits characteristically linked to anhedonia in clinical literature. Overall, the results provide a baseline for depression detection in Modern Greek and underscore the importance of grounding automated screening in clinically interpretable evidence.

Keywords: depression detection, social media, mental health

1. Introduction

Natural Language Processing (NLP) research plays a pivotal role in advancing mental health disorders such as depression screening through the development of sophisticated speech and text based models (Gómez-Zaragozá et al., 2025; Liu et al., 2022). To date, research has predominantly utilized social media data (De Choudhury et al., 2013; Eichstaedt et al., 2018), with a heavy reliance on platforms like X (formerly Twitter) and Reddit (Harrigian et al., 2021). While the majority of these studies focus on English-language resources, recent efforts have begun to bridge the gap for other languages, including Portuguese (Santos et al., 2023), German (Zanwar et al., 2023), Arabic (Almouzini et al., 2019), and Chinese (Zhang et al., 2024).

This shift toward linguistic diversity is exemplified by recent efforts to consolidate multilingual research, such as the first comprehensive survey of mental disorder detection in non-English languages (Bucur et al., 2025). Exploring diverse languages is essential to determine whether linguistic cues of depression are universal or subject to cultural and contextual variation.

Notwithstanding this progress, two significant gaps remain in the literature: (i) methodological opacity; studies utilize Machine Learning (ML) or Deep Learning (DL) architectures to classify depressed vs. non-depressed individuals (Tadesse et al., 2019; Husseini Orabi et al., 2018) which lack interpretability required for clinical confidence and (ii) lack of explainability; existing methods fail to provide the reasoning behind a classification. In mental health contexts, it is crucial to understand how automated markers correlate with established clinical symptomatology (Zhang et al., 2022).

In this work, we present a comprehensive evaluation of text-based models using a social media dataset in Modern Greek (MG). Our study establishes a performance baseline for depression screening in an under-resourced language, contributing to the broader goal of linguistic inclusivity in NLP. Furthermore, we leverage Linguistic Inquiry and Word Count (LIWC) features (Pennebaker et al., 2015) to enhance model interpretability. By mapping specific linguistic patterns to classification outcomes, we investigate the extent to which these cues align with documented clinical symptoms of depression.

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the Modern Greek datasets utilized in this study. Section 3 outlines the experimental framework, detailing the preprocessing pipeline and the selection of both traditional baselines and transformer-based models. In Section 4, we present the experimental results and analysis. Section 5 summarizes our findings and suggesting avenues for future research in multilingual mental health NLP.

Proceedings of the 6th RaPID@MENTAl.ai Workshop @ LREC 2026, pages 106–114
12 May 2026.

2. Dataset description

Previous research highlights the effectiveness of self-disclosure statements, where users are identified based on explicit mentions of a depression diagnosis, as reliable proxies for depression detection in social media corpora (Jagfeld et al., 2021; Jamil et al., 2017; Coppersmith et al., 2015). Following these established practices, our study utilizes the Modern Greek depression and control corpora compiled by Stamou et al. (2024).

In particular, the Depression Corpus (DC) was developed by isolating tweets containing explicit self-reports, specifically: 'I have been diagnosed with depression. The dataset includes 51 unique users, whose profiles underwent manual filtration to exclude instances of humor, off-topic references, and automated health news feeds. Subsequently, a comprehensive corpus of 659,189 tweets was constructed by retrieving the longitudinal tweet history for these individuals.

In addition to this, the work utilizes two distinct control datasets. The first, Control Corpus 1 (CC1), follows the framework of Chancellor and Choudhury (2020) and consists of non-depressed users identified through random sampling. This dataset was restricted to Greek-language users with no documented history of mental health disorders (i.e. no reference to mental health issues). The second, Control Corpus 2 (CC2), is a topic derived corpus to ensure that the control group reflects the same thematic concerns typically expressed by users in the depression group. By matching the control data to these specific areas of interest, the authors aimed to minimize thematic bias and ensure that classification relies on linguistic markers of depression rather than mere differences in subject matter.

This combination of control corpora provided a reliable, balanced foundation for subsequent depression detection experiments. From the initial pool of 659,189 tweets, we extracted randomly a subset of 10K posts for the modeling phase. This subsampling strategy was employed to ensure a balanced distribution across the depression and control classes, preventing model bias toward high-volume users while maintaining computational efficiency.

3. Experimental setup

3.1. Data preprocessing

The primary objective of this study was to develop a binary classifier capable of distinguishing between depressed and non-depressed individuals. To ensure high data quality before model training, the corpora were preprocessed by removing URLs and normalizing repetitive punctuation (e.g., "!!!!"). We utilized the Stanza library (Qi et al., 2020) for robust tokenization and performed stopword removal to reduce feature noise. Furthermore, we excluded all tweets containing fewer than two words to ensure sufficient linguistic context for the models.

3.2. Baseline

For the experiments, we employed PyCaret (Ali, 2020), an open-source tool that provides several ML classifiers. This allowed for the systematic benchmarking and comparison of twelve distinct machine learning classifiers, leveraging its integrated modules for model selection and performance evaluation.

- Linear & Statistical: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), SVM (Linear Kernel), and Naive Bayes (NB)
- Tree-Based: Random Forest (RF), Extra Trees (ET), AdaBoost, and Decision Trees (DT)
- Distance Based: K-Neighbors (KNN) and Multi-layer Perceptron (MLP)
- Baseline: A Dummy Classifier was included for performance comparison.

Text was encoded using two distinct feature extraction approaches:

- TF-IDF vectorization: We used an n-gram range of (1, 3) with a maximum of 5,000 features. This approach mitigates the dominance of high-frequency words by calculating the Term Frequency-Inverse Document Frequency, thereby amplifying the significance of rare, semantically rich terms.
- LIWC Lexicon: Features were extracted using the Linguistic Inquiry and Word Count (LIWC) tool. Unlike the TF-IDF approach, punctuation and accentuation were retained here, as the LIWC dictionary utilizes these markers to capture psychological and emotional states. In Greek, removing diacritics may merge distinct lexical forms, potentially affecting LIWC category assignments. Additionally, punctuation marks (e.g., exclamation and question marks) contribute to the detection of affective intensity and cognitive processes, which are central to LIWC-based analysis.

Experiments were conducted using an 80/20 train-test split with 5-fold cross-validation. Dataset splitting was performed at the user level rather than the tweet level. Specifically, tweets were grouped by unique user IDs prior to splitting, ensuring that all tweets from a given user appear exclusively in

either the training or test set. All feature variables were scaled using Z-score normalization to ensure a mean of 0 and a standard deviation of 1.

3.3. Transformer-based models

We evaluate several transformer-based models to assess the impact of different pretraining corpora on depression detection in MG:

- `bert-base-greek-uncased-v1`: A BERT model specifically trained on Modern Greek corpora.
- `greeksocialbert-base-v2`: A model pretrained on Greek social media data, potentially capturing platform-specific nuances.
- `bert-base-multilingual-cased`: A multilingual model trained on 104 languages, including Greek.
- `xlm-roberta-base`: A robust multilingual model optimized for cross-lingual transfer.

All models were fine-tuned for 4 epochs with a learning rate of 10^{-5} and a batch size of 16. The experiments were executed in a Google Colaboratory environment equipped with 12 GB of Virtual Memory and an NVIDIA Tesla T4 GPU. For the machine learning baselines described in Section 3.2, all feature variables (TF-IDF and LIWC) were scaled using Z-score normalization to ensure a mean of 0 and a standard deviation of 1. To ensure the reproducibility of our results, the session seed was fixed to `123`.

4. Results

This section presents the experimental results on depression detection and their analysis.

4.1. Baseline

Evaluation on the CC2 We conducted the experiments on a a balanced sample of the corpus, consisting of 10K tweets in total (5,000 from the depression corpus and 5,000 from the CC2 control group). As illustrated in Table 1, we observed that increasing the training set size beyond this point yielded diminishing returns. Doubling the corpus to 40K tweets resulted in only marginal gains in the F1-score, suggesting that 10K tweets provide a sufficient representation of the underlying linguistic patterns for this task.

Table 2 summarizes the performance of twelve classifiers using LIWC features. The evaluation includes standard metrics such as Accuracy, Precision, Recall, and F1-score, along with the Kappa statistic, which assesses the agreement between predicted and true labels while accounting for chance. Additionally, the Matthews Correlation Coefficient (MCC), a robust measure that considers true and false positives and negatives, is presented.

Train Size	Acc	Prec	Rec	F1	F_{mac}
10k	0.9400	0.9434	0.9367	0.9398	0.9399
20k	0.9534	0.9498	0.9575	0.9536	0.9534
40k	0.9576	0.9453	0.9716	0.9582	0.9576

Table 1: Performance as a function of training set size.

At first glance, based on the Accuracy metric, the top-performing models are LightGBM (F1 = 0.7123) and Extra Trees (F1 = 0.6779). However, a closer look reveals a small imbalance between precision and recall across these models. Beyond predictive power, a primary goal of this study is to ensure model interpretability. Tree-based ensembles are particularly suited for this, as they allow for the extraction of feature importance and decision paths. Consequently, we selected the Extra Trees Classifier ('et') as our primary model for further analysis, as it offers a superior balance between competitive performance and the transparency required for LIWC feature interpretation.

Table 2: Classification results with LIWC features on the CC2.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
LightGBM	**0.7170**	**0.7938**	0.7014	0.7238	0.7123	**0.4339**	0.4343
Extra Trees	0.7128	0.7759	0.6055	**0.7706**	0.6779	0.4256	**0.4358**
Random Forest	0.7125	0.7807	0.6122	0.7652	0.6801	0.4248	0.4336
Grad. Boosting	0.7125	0.7873	0.7134	0.7117	0.7125	0.4249	0.4250
AdaBoost	0.7086	0.7792	0.7262	0.7011	**0.7134**	0.4172	0.4175
KNN	0.6555	0.7148	0.7480	0.6309	0.6844	0.3112	0.3168
Decision Tree	0.6461	0.6157	0.6693	0.6392	0.6539	0.2923	0.2927
QDA	0.6161	0.7437	0.8449	0.5794	0.6874	0.2326	0.2616
Naive Bayes	0.6000	0.7452	**0.8630**	0.5652	0.6831	0.2004	0.2356
Linear SVM	0.5782	0.5578	0.7244	0.5571	0.6263	0.1566	0.1696
Ridge	0.5450	0.6182	0.5604	0.5432	0.5516	0.0901	0.0902
LDA	0.5450	0.6182	0.5604	0.5432	0.5516	0.0901	0.0902
Logistic Reg.	0.5440	0.6151	0.5576	0.5423	0.5498	0.0881	0.0882
Dummy	0.5005	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Interestingly, when switching to TF-IDF features (Table 3), performance improved significantly across all models, with Ridge and LDA exceeding 0.87 in accuracy. This suggests that while LIWC captures psychological markers, the broader lexical variety captured by TF-IDF provides stronger discriminative signals for this task.

4.2. Model interpretability and feature analysis

Building on the selection of the Extra Trees Classifier for its balance of performance and transparency, we conducted a feature attribution analysis to identify the linguistic markers of depression. To better understand which linguistic features most influence predictions, we analyzed SHAP (SHapley Additive

Table 3: Classification results with TFIDF features on the CC2.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
Ridge	**0.8783**	0.0000	0.8877	0.8713	0.8794	0.7566	0.7567
LDA	**0.8783**	0.9433	0.8877	0.8714	**0.8795**	**0.7566**	**0.7568**
LightGBM	0.8741	**0.9437**	0.8791	0.8704	0.8747	0.7482	0.7483
Extra Trees	0.8721	0.9430	0.8239	**0.9118**	0.8656	0.7442	0.7477
Random Forest	0.8430	0.9232	0.7890	0.8847	0.8341	0.6861	0.6902
Linear SVM	0.8371	0.0000	0.8283	0.8431	0.8356	0.6741	0.6743
Logistic Reg.	0.8354	0.8956	0.8461	0.8284	0.8371	0.6708	0.6710
AdaBoost	0.8021	0.8785	0.8553	0.7731	0.8121	0.6042	0.6077
Grad. Boosting	0.7979	0.8807	0.8347	0.7775	0.8051	0.5958	0.5975
Naive Bayes	0.7860	0.7877	**0.9140**	0.7277	0.8103	0.5720	0.5917
Decision Tree	0.7848	0.7767	0.7773	0.7891	0.7832	0.5696	0.5697
KNN	0.6048	0.6535	0.8916	0.5666	0.6928	0.2096	0.2559
QDA	0.5739	0.5739	0.1542	0.9597	0.2657	0.1477	0.2717
Dummy	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

exPlanations) values and present a beeswarm plot in Figure 1.

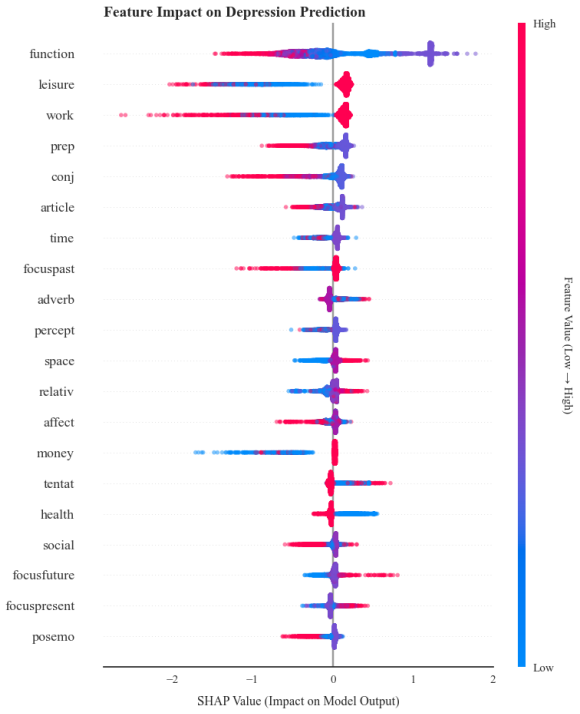


Figure 1: SHAP summary plot showing the top LIWC 20 features contributing to the prediction of the depressed class.

The beeswarm graph can be interpreted as follows. The visualization shows the SHAP values for every feature for every instance in each dataset, where each dot represents a SHAP value for a specific feature. The x-axis indicates the SHAP value magnitude, and the y-axis lists the features. In addition, the color of the dot reveals the value of the feature for that instance (e.g., red for high, blue for low). The features are ordered by the mean SHAP values. They are presented in a descending order of significance along the y-axis. A dot positioned on the right-hand side of the graph signifies a positive impact, while one on the left side indicates a negative impact. For instance, the LIWC-category 'work' showcases notably high negative SHAP values. The position of a data point along the x-axis signifies the degree of influence it exerts. Therefore, the further away from the point of origin (0), the greater the impact it has on the model's predictions.

The analysis reveals several critical insights:

- **Reduced engagement:** The LIWC categories `work`, `leisure`, and `social` exhibit high negative SHAP values. This indicates that a high frequency of words related to these domains is strongly associated with the non-depressed class. Clinically, this aligns with the social withdrawal and reduced interest in activities typical of depressive episodes.

- **Loss of drive and anhedonia:** The LIWC category `drives`, which include aspirations, achievements, and rewards, do not appear as prominent markers for the depressed class. This mirrors the clinical symptom of **anhedonia**, where individuals experience lowered expectations of reward and impaired reinforcement learning (Barch et al., 2016).

- **Temporal focus:** The `past tense` feature exerts a negative influence on the model's identification of the depressed class. This suggests that depressed individuals may focus less on past experiences and more on their current emotional state.

Considering that depression can lead to a lack of motivation, reduced interest in activities, and a sense of hopelessness, it is normal to expect that individuals face difficulties in pursuing their ambitions (Watson et al., 2020). Furthermore, the 'drives' LIWC category in LIWC 2015 version, includes also the 'achievements' category, which refers to experiences of success, and also the 'reward' category, which relates to the experience of receiving positive reinforcement. Many studies support that a core symptom of depression, usually referred as "anhedonia" is associated with lowered expectations of rewards and impaired reinforcement learning (Barch et al., 2016; Treadway and Zald, 2011).

Evaluation on the CC1 A second evaluation setting involved a comparison between the selected subset of the depression corpus and the randomly sampled control corpus. The subset from the depression corpus corresponded to a single user who served as a gold reference case. In particular, this user corresponded to an individual with a confirmed severe outcome, which was treated as a benchmark instance to approximate a real-world scenario of extreme depression risk. This setting should be interpreted as a stress-test scenario rather than a

general population model. By contrasting a confirmed severe case against a highly similar control user, we aim to evaluate whether linguistic signals remain detectable even under strong lexical similarity constraints.

To enable a fair comparison, we introduced a similarity-based control selection procedure. Cosine similarity was computed between vector representations of individual users and the gold reference user. Word embeddings were generated using the Greek FastText model [1]. The most similar non-depressed user achieved a similarity score of 0.99 but contained only 58 tweets, which was insufficient for reliable comparison. Therefore, the second most similar user (similarity = 0.97), with 12,076 tweets, was selected.

Tweets were preprocessed by removing URLs, usernames, punctuation, and emojis (via the demoji package), and by retaining only tweets containing more than two words. A random subsample of 10,000 tweets was selected to match previous experimental settings.

LIWC-based features Results using LIWC features (Table 4) show low classification performance, with accuracy ranging between 50% and 58%, only marginally above chance level. Most models achieved and MCC values close to zero, further indicating weak discriminative power. This behavior is attributed to the sparse coverage of the LIWC lexicon in the Greek Twitter corpus, resulting in sparse feature representations. Consequently, LIWC-based results were not further analyzed.

TF-IDF-based features TF-IDF features were then employed (Table 5). In contrast to LIWC, TF-IDF representations substantially improved performance across models. The Extra Trees classifier achieved the best overall results (Accuracy = 0.7150, F1 = 0.7332, = 0.4300, MCC = 0.4340), corresponding to an improvement of approximately 10 percentage points in accuracy compared to LIWC-based models.

Although TF-IDF features significantly outperform LIWC, several models exhibit noticeable discrepancies between Precision and Recall, suggesting instability in class-wise behavior. In particular, some classifiers favor recall at the expense of precision, indicating potential class bias. Further investigation is required to better understand the factors contributing to this imbalance.

4.3. Transformer-based models

We evaluated the performance of four pretrained models: (i) `bert-base-greek-uncased-v1`, (ii) `greeksocialbert-base-v2`, (iii) `bert-base-multilingual-cased` (mBERT), and (iv) `xlm-roberta-base`. This evaluation focuses on the CC2 dataset, which utilizes a control group constructed via topic-oriented extraction to ensure thematic consistency. The comparative metrics are summarized in Table 4.3.

[1] https://huggingface.co/facebook/fasttext-el-vectors

Table 4: Model performance with LIWC features; gold depressed vs random control user.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
KNN	**0.5865**	0.6103	0.5858	0.5867	0.5862	**0.1730**	**0.1730**
Logistic Reg.	0.5801	**0.6187**	0.5977	0.5795	0.5851	0.1602	0.1625
MLP	0.5788	0.6178	0.6477	0.5719	**0.6043**	0.1575	0.1610
Extra Trees	0.5511	0.5743	0.5872	0.5556	0.5616	0.1022	0.1065
Linear SVM	0.5432	0.0000	0.4612	0.5533	0.4957	0.0865	0.0894
LDA	0.5137	0.5236	**0.7192**	0.5099	0.5965	0.0275	0.0297
Decision Tree	0.5076	0.5168	0.3295	0.5898	0.3197	0.0153	0.0225
AdaBoost	0.5066	0.5127	0.6022	0.5298	0.4076	0.0133	0.0649
Random Forest	0.5037	0.5636	0.6222	0.4899	0.4401	0.0075	0.0463
Naive Bayes	0.5024	0.5992	0.0092	0.6242	0.0182	0.0047	0.0238
QDA	0.5022	0.6000	0.0082	0.6368	0.0162	0.0045	0.0243
Dummy	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 5: Model performance with TFIDF features; gold depressed vs randomly selected user.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
Extra Trees	**0.7150**	0.7804	0.7831	0.6894	**0.7332**	**0.4300**	**0.4340**
MLP	0.7113	**0.7824**	0.7581	0.6933	0.7240	0.4225	0.4247
Random Forest	0.6981	0.7722	0.7969	0.6656	0.7252	0.3962	0.4044
Linear SVM	0.6880	0.0000	0.7744	0.6605	0.7125	0.3759	0.3823
Logistic Reg.	0.6850	0.7450	0.6909	0.6830	0.6868	0.3700	0.3702
Naive Bayes	0.6764	0.6838	0.5772	**0.7200**	0.6406	0.3528	0.3600
Decision Tree	0.6566	0.6612	0.7084	0.6416	0.6733	0.3131	0.3150
AdaBoost	0.6375	0.7077	**0.9216**	0.5877	0.7176	0.2750	0.3349
KNN	0.6175	0.6480	0.7181	0.5982	0.6516	0.2350	0.2416
LDA	0.5986	0.6124	0.5944	0.5991	0.5966	0.1972	0.1973
QDA	0.5456	0.5456	0.7856	0.5324	0.6285	0.0913	0.1108
Dummy	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Model	Acc.	Prec.	Rec.	F1
Greek BERT	0.948	0.938	**0.960**	0.949
GreekSocialBERT	0.942	0.934	0.953	0.942
mBERT	**0.953**	**0.965**	0.940	**0.952**
XLM-Roberta Base	0.944	0.961	0.926	0.943

Table 6: Model performance on the CC2.

Evaluation on the CC2 All transformer models achieved strong performance, with accuracy ranging from 94.2% to 95.3%. The multilingual BERT model obtained the highest overall accuracy (0.953), while Greek BERT achieved slightly higher recall (0.960), indicating strong sensitivity to depressed instances.

For Greek BERT, recall exceeds precision by approximately 2.2 percentage points (0.960 vs 0.938). In the context of depression detection, prioritizing recall may be desirable, as minimizing false negatives (i.e., failing to identify depressed individuals) is often considered critical. In contrast, multilingual BERT achieves higher precision (0.965) but slightly lower recall (0.940), suggesting a more conservative classification strategy. This model produces fewer false positives but may miss a greater num-

ber of true depression cases compared to Greek BERT.

Overall, these transformer architectures substantially outperform the traditional feature-based baselines presented in Section 4.1. This performance leap demonstrates the effectiveness of contextualized embeddings in capturing the nuanced semantic and syntactic structures associated with mental health discourse.

Evaluation on the CC1 We next evaluated the transformer models under the second experimental setting, comparing the gold depressed user with the cosine-similarity-matched control user (Table 4.3).

Performance decreases notably in this constrained setting, with accuracy ranging from 0.554 (`xlm-roberta-base`) to 0.779 (`greeksocialbert-base-v2`). GreekSocialBERT achieved the best overall performance (F1 = 0.776), followed by mBERT (F1 = 0.758). While mBERT showed a slightly more balanced precision-recall trade-off, GreekSocialBERT's superior scores suggest that its pretraining on social media data may provide an advantage when distinguishing subtle linguistic nuances in informal Greek text.

Model	Acc.	Prec.	Rec.	F1
Greek BERT	0.753	0.766	0.736	0.748
GreekSocialBERT	**0.779**	**0.788**	**0.767**	**0.776**
mBERT	0.759	0.761	0.758	0.758
XLM-Roberta base	0.554	0.717	0.534	0.420

Table 7: Model Performance on the CC1.

The significant performance drop compared to the CC2 experiment is expected and directly attributable to the similarity-based sampling procedure. Because the control user was selected based on extreme cosine similarity (0.97), the lexical and distributional overlap between the two classes is substantial. This minimizes "topic-driven" separability, where a model might simply distinguish between "talking about sadness" vs. "talking about sports", and forces the model to rely on much more granular linguistic markers.

This experiment serves as a stress-test scenario, evaluating whether models can maintain discriminative power when surface-level similarity is high. The overlap in vector space likely causes the transformer models to struggle with defining clear decision boundaries, as the embeddings for both classes occupy nearly identical regions

We avoid attributing the performance drop solely to overfitting, as the primary cause appears to be reduced inter-class variance due to similarity constraints. Moreover, while transformer models achieve strong predictive performance, they offer limited interpretability compared to feature-based approaches. Because classification decisions are based on high-dimensional contextual representations, isolating specific linguistic markers that distinguish depressed from non-depressed users remains challenging.

Moreover, while transformer models offer superior predictive power, this experiment highlights a critical trade-off: performance vs. interpretability. Unlike the feature-based analysis (LIWC/SHAP), the decision-making process of these high-dimensional contextual models remains opaque. While their performance is better in comparison to traditional models, they cannot explicitly reveal which Greek linguistic cues were the deciding factors.

5. Conclusions

We evaluated a range of machine learning (ML) and deep learning (DL) architectures for the binary classification of depressed versus non-depressed individuals. Our results identify mBERT as the overall top-performing model, achieving an accuracy and F1-score of 0.95 on the topic-oriented corpus (CC2). However, performance decreases substantially in the stress-test evaluation (CC1), where models are required to distinguish a gold-standard depressed user from a linguistically similar control user. This drop is likely due to the increased difficulty of the task, as the control user was selected based on high cosine similarity (0.97) to the target user. This high degree of lexical and semantic similarity may reduce class separability and increase classification ambiguity, suggesting that surface-level representations may be insufficient in highly controlled matching scenarios and that more fine-grained linguistic features may be beneficial.

For the ML baselines, we contrasted TF-IDF and LIWC features. TF-IDF achieved the highest predictive performance, with the Extra Trees (ET) classifier reaching an accuracy of 0.86 and an F1-score of 0.85. However, LIWC features offered greater interpretability, enabling direct mapping of linguistic patterns to meaningful LIWC categories.

Despite the more moderate performance of LIWC-based models (e.g., ET achieving 0.65), we utilized the ET classifier for our primary feature analysis due to its inherent robustness and the interpretability of its decision paths. LIWC features as a part of the feature engineering component has been exploited in numerous studies (Coppersmith et al., 2015; Resnik et al., 2013; Asgari et al., 2017; Tadesse et al., 2019). Their incorporation alongside other features has consistently demonstrated their potential to enhance classification performance, underscoring their significant role in capturing language-specific cues related to

depression. In this study, the interpretability analysis identified several prominent linguistic markers within the Greek depression corpus: (i) social withdrawal, manifested as a marked decrease in discourse related to social and leisure activities; (ii) professional disengagement, characterized by a significant reduction in work-related engagement; (iii) diminished drive, reflected in lower motivational cues and fewer references to achievements or rewards; and (iv) altered temporal orientation, evidenced by shifting patterns in time-reference words relative to the control group. These findings align with established clinical literature and support the existence of language-specific cues in depressive communication.

Ultimately, this study advocates for grounding automated screening in clinically validated evidence. While digital self-disclosure offers a useful proxy for risk, verifying these patterns against formal diagnoses remains a priority for future work. Establishing a clinically validated “ground truth” is essential to ensure that detected pattern, such as anhedonia, social withdrawal, and reduced engagement, accurately reflect the underlying depressive condition.

6. Copyrights

7. Acknowledgements

This work is part of the first author’s doctoral thesis. «The implementation of the doctoral thesis was cofinanced by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the Act “Enhancing Human Resources Research Potential by undertaking a Doctoral Research” Subaction 2: IKY Scholarship Programme for PhD candidates in the Greek Universities».

8. Ethical considerations

The use of social media data for mental health screening raises significant ethical challenges regarding privacy, data ownership, and potential harm. Our study adheres to the following ethical principles: (i) data privacy and anonymization: Although the data used in this study (CC1 and CC2) were collected from public social media platforms, we recognize that users may not have intended for their posts to be used in a psychiatric research context. To protect user identity, all personal identifiers (i.e., usernames, locations etc.) were removed, (ii) this research is strictly observational, we did not engage with any users during the data collection process, and (iii) the models are intended as screening aids for researchers and clinicians, not as definitive diagnostic tools.

Beyond privacy, additional ethical concerns arise from the potential deployment of such systems. In particular, diagnostic errors, such as false positives and negatives, may lead to unintended consequences, including unnecessary stigmatization of users or failure to identify individuals who may require support.

9. Limitations and Future Work

Despite the contributions of this work, several limitations should be considered when interpreting the results:

1. Platform bias and corpus representativeness: our dataset is derived exclusively from social media (X/Twitter). These platforms tend to skew toward specific demographics, often younger individuals, which may not fully represent the linguistic patterns of the broader speaking population or those with different socio-economic backgrounds.

2. "Silver standard": we acknowledge that labels are based on self-reported diagnoses. While these provide a "silver standard" for training, they lack the clinical precision of a formal psychiatric evaluation conducted by a mental health professional.

3. Linguistic resource constraints: While we utilize LIWC to bridge the explainability gap, it is important to note that the Greek version of the LIWC dictionary, while robust, may not capture the full nuanced range of informal "internet slang" in the Greek digital landscape as effectively as the original English version.

4. Stress-test generalizability: The stress-test experiment, which includes a single user in the

depression class, may capture idiolectal (user-specific) linguistic patterns rather than generalizable markers of depression. As such, its results should be interpreted as exploratory.

Future work will extend explainability to transformer-based models using methods such as SHAP or integrated gradients, to better understand which features contribute to depression prediction.

10. Bibliographical References

Moez Ali. 2020. *PyCaret: An open source, low-code machine learning library in Python*. PyCaret version 1.0.0.

Salma Almouzini, Maher khemakhem, and Asem Alageel. 2019. Detecting arabic depressed users from twitter data. *Procedia Comput. Sci.*, 163(C):257–265.

Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. 2017. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228.

Deanna M. Barch, David Pagliaccio, and Katherine R. Luking. 2016. Mechanisms underlying motivational deficits in psychopathology: similarities and differences in depression and schizophrenia. *Current Topics in Behavioral Neurosciences*, 27:411–449.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2025. A survey on multilingual mental disorders detection from social media data. *CoRR*, abs/2505.15556.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, 3.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of ICWSM*, pages 128–137.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preotiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115:11203–11208.

Lucía Gómez-Zaragozá, Javier Marín-Morales, Mariano Alcañiz, and Mohammad Soleymani. 2025. Speech and text foundation models for depression detection: Cross-task and cross-language evaluation. *Interspeech 2025*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Glorianna Jagfeld, Fiona Lobban, Paul Rayson, and Steven Jones. 2021. Understanding who uses Reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 1–14, Online. Association for Computational Linguistics.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, and Jing Guo. 2022. Detecting and measuring depression on social media using a machine learning approach: Systematic review. *JMIR Ment Health*, 9(3):e27244.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates, Austin, TX.

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.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, Washington, USA. Association for Computational Linguistics.

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2023. Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Lang. Resour. Eval.*, 58(1):273–300.

Vivian Stamou, George Mikros, George Markopoulos, and Spyridoula Varlokosta. 2024. Establishing control corpora for depression detection in Modern Greek: Methodological insights. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 68–76, Torino, Italia. ELRA and ICCL.

Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Michael T. Treadway and David H. Zald. 2011. Reconsidering anhedonia in depression: lessons from translational neuroscience. *Neuroscience and biobehavioral reviews*, 35 3:537–55.

Rebecca Watson, Kate Harvey, Ciara McCabe, and Shirley Reynolds. 2020. Understanding anhedonia: A qualitative study exploring loss of interest and pleasure in adolescent depression. *European Child & Adolescent Psychiatry*, 29(4):489–499.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. SMHD-GER: A large-scale benchmark dataset for automatic mental health detection from social media in German. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1526–1541, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhenwen Zhang, Jianghong Zhu, Zhihua Guo, Yu Zhang, Zepeng Li, and Bin Hu. 2024. Natural language processing for depression prediction on sina weibo: Method study and analysis. *JMIR Mental Health*, 11.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Profiling Psychopathic Behavior Using Machine Learning

Avraham Treistman, Tehilla David, Sivan Levi, Dror Mughaz
Jerusalem COllege of Technology
Department of Computer Science, Jerusalem, Israel, 9372115 , Address2, Address3
treistma@jct.ac.il, sivanlevi94@gmail.com, tehila859@gmail.com, myghaz@gmail.com
{Avraham Treistman, Tehilla David, Sivan Levi, Dror Mughaz

Abstract

Psychopathy is a complex personality disorder characterized by persistent deficits in empathy and manipulative behavior. Traditional diagnostic methods often rely on subjective clinical assessments, which are susceptible to deception. This research proposes an objective, non-invasive computational framework for profiling psychopathic traits using Natural Language Processing (NLP) and Machine Learning. We developed a systematic pipeline utilizing transcribed interviews from confirmed criminal psychopaths and a balanced control group. To address data sparsity and noise, we employed the Dynamic Variance Thresholding (DyVaT) algorithm to construct a semantically dense vocabulary of over 1,300 features. The methodology integrates advanced preprocessing, TF-IDF vectorization, and synonym-based data augmentation to enhance model generalization. Among the evaluated classifiers, a Linear Support Vector Machine (SVM) achieved the highest performance, with an accuracy of 0.8081 and an F1-score of 0.7957. Our findings demonstrate the efficacy of linguistic biomarkers and feature importance analysis in distinguishing psychopathic speech patterns. This study provides a scalable methodology for early screening and diagnostics, with significant implications for forensic psychology, security, and ethical AI deployment in mental health.

Keywords: Computational Linguistics, Machine Learning Application, Psychopathy Profiling, Linguistic Biomarkers, Data Augmentation, Dynamic Variance Thresholding

1. Introduction

Psychopathy is a severe personality disorder with major social and clinical consequences. It involves low empathy, manipulation, and persistent antisocial behavior. Many individuals show a convincing "mask of sanity" in everyday interactions. The diagnosis often relies on the PCL-R checklist plus lengthy face-to-face interviews. These procedures are difficult to scale and can vary between clinicians. They also depend on cooperation, which deceptive subjects may deliberately undermine.

We therefore need complementary screening methods that are more objective and auditable. Computational pipelines can support clinicians without replacing clinical judgment. They fit remote data collection and monitoring workflows that are increasingly realistic. They also encourage shared resources, clear protocols, and repeatable evaluation. Still, the tools must be interpretable, safe, and cautious in high-stakes use. Ethics, consent, privacy, and bias management must be designed from the beginning.

NLP is promising because language leaks subtle patterns during spontaneous narration. Speakers rarely control these patterns consistently throughout an interview. Machine learning has identified "Dark Triad" traits from text (Yeasmin et al., 2024). Psychopathy detection remains difficult because labeled, verified transcripts are scarce. Legal barriers and sensitivity restrict access, creating a severe class imbalance. Domain shifts across sources can also distort models unless explicitly handled.

Our study presents a pipeline designed around these practical data constraints. We focus on distinguishing criminal psychopaths from a non-psychopathic control group. The offender interviews were collected from YouTube, while the controls came from NPR interviews. We matched conversational style and cleaned transcripts using consistent annotation rules. This reduced cues from formatting, editing, or interviewer structure. It also supports reproducibility and future sharing of datasets and tools.

Representation choices were critical, so we avoided using every observed token. Instead, we built a seed list of psychologically relevant words by manual selection. We expanded that list using Dynamic Variance Thresholding, or DyVaT (Treistman et al., 2022). DyVaT retains semantically related terms while filtering high-variance noise from embeddings. The final lexicon contains about 1,357 focussed features for modeling. This feature design improves interpretability and helps address sparsity and dimensionality issues.

We then applied a rigorous text-processing workflow to ensure consistent input. Transcripts were segmented into standardized 15-sentence chunks to stabilize sample length. Lemmatization reduced sparsity by mapping inflected forms to base roots. Data scarcity persisted, so we augmented the minority class with synonym substitution. Such augmentation is common for imbalance in personality classification (Pradana and Suhartono, 2024). We

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 115–125
12 May 2026.

treated augmentation conservatively, since it can introduce small semantic drift.

For classification, we compared Logistic Regression, Random Forest, and Support Vector Machines. Choosing the right classifier is central in text categorization (Allam et al., 2025). Random Forest often handles high-dimensional data well (Venkateshwarlu et al., 2024). However, our best results came from a Linear SVM in this setting. It achieved 80.81% accuracy and a 0.7957 F1 score (Alzoubi et al., 2023). We also tracked feature significance to support transparent screening and diagnostics.

Feature inspection revealed systematic differences that align with psycholinguistic expectations. Psychopathic speech overused basic-need terms like “money,” “food,” and “sex.” It also used more words related to violence, dominance, and authority. Controls used more language about social connection, norms, and morality (Adkins et al., 2025). High-stakes use demands transparency, consent, privacy safeguards, and bias checks (Zhou and Chen, 2023). Next, we will test LLMs, acoustic cues, and sensor-ready multimodal designs on lightweight assessment platforms.

2. Related Works

Classic machine learning remains competitive for high-dimensional text features. Random Forest often performs well in sparse spaces (Venkateshwarlu et al., 2024). Support Vector Machines can be robust across languages and settings (Alzoubi et al., 2023). A survey of these methods and common evaluation practices guides model selection under real deployment constraints (Gasparetto et al., 2022).

Work on personality and “Dark Triad” traits motivates psychopathy-oriented screening . Standard classifiers can separate higher-risk profiles from controls (Yeasmin et al., 2024). There are reported gains from ensembles on non-linear personality signals (Maxim et al., 2025). Regex rules can be mixed with NLP for phase-aware disorder detection (Patel and Johnson, 2025). Research connects deception and emotion signals to forensic text analysis (Adkins et al., 2025).

Data collection and labeling protocols strongly shape what models actually learn. Many studies depend on secondary data, which complicates consent and reuse. Shared schemas, annotation rules, and audit trails help build usable infrastructure. These steps also support domain adaptation when sources differ in style.

Preprocessing decisions are not neutral in psychological profiling tasks. Preprocessing can change accuracy, even for sentiment pipelines (Alam and Yao, 2019). Stop word removal is especially tricky for self-reference and social framing (Kaur and Buttar, 2018), and psychological settings need extra caution. Tokenization quality also matters for mental health signals (Dixit et al., 2023).

Feature weighting and representation are still common baselines in clinical text modeling. Various TF-IDF variants are compared for use on unstructured datasets, such as interview transcripts (Das et al., 2023). However, “noise” can contain the most diagnostic information linking psychopathic deviation to narrative style and lexical choices (Gawda, 2022).

A central innovation of our work is the method of vocabulary construction. We did not simply use all the words available in the transcripts. Instead, we used a "seed list" of manually selected words that are psychologically relevant. We expanded this list using the Dynamic Variance Thresholding (DyVaT) algorithm (Treistman et al., 2022). This algorithm helps identify semantically relevant words while filtering out high-variance noise. The result is a focused lexicon of approximately 1,357 features. This approach allows us to target the specific narrative structures associated with psychopathy. Deep learning is increasingly used to model subtle and contextual personality signals. A review of machine and deep learning for trait detection notes the growing use of ensembles and larger architectures (Naz et al., 2025). Graph approaches can model relations between words, users, and contexts, such as LL4G for self-supervised, dynamic graph-based personality detection (Shen et al., 2025).

Network science offers another angle on discourse and psychological structure, such as textual forma mentis networks for adolescents and psychopathology levels (Carrillo et al., 2025). Knowledge-guided filtering can focus models on clinically informative segments such as PsyTEx for refining text for psychological analysis (Bhandarkar et al., 2025). Emotion dynamics can serve as biosocial markers beyond static sentiment (Teodorescu et al., 2023).

Related mental health domains also shape methods we can reuse or adapt, such as enhanced TextGCN for depression detection using emotion representations (Mao and Han, 2025) or DepGLM to recognize degrees of depression with LLM support (Liu et al., 2025). These tasks differ from psychopathy, but the modeling patterns are transferred. They also raise similar needs for calibration and clinically meaningful metrics. The scarcity of labeled psychopathy data remains a practical bottleneck for supervised learning. The class imbalance is severe, and verified labels are rarely available to share. This imbalance can be addressed by augmentation, such as intent-aware (Saleem and Kim, 2024) or synonym replacement(Pradana and Suhartono, 2024).

Other pipelines generate synthetic samples for downstream interventions and support tools, such as ERNIE-based augmentation for CBT-related applications (Sambana et al., 2025). Topic modeling plus synthetic generation improved suicidal ideation detection (Ghanadian et al., 2025). These results suggest augmentation can help generalization when clinical data is constrained. Still, synthetic text can drift and must be validated carefully. Large language models are changing baselines, but introduce new risks, such as the applicability of LLMs for the classification of health text using public social media data (Guo et al., 2024), whether GPT-3 exhibits psychopathic traits under psychological perspectives (Li et al., 2022), questioning whether human personality tests can be applied to algorithmic agents (Sühr et al., 2025), or LLM blurring of linguistic markers used for trait inference (Sourati et al., 2024).

Safety and governance are central when models touch sensitive psychological labels (Li et al., 2024). Ethical principles must be applied to engineering practice (Zhou and Chen, 2023; Mittelstadt, 2019) while addressing regulatory gaps in AI-driven profiling on social media (Bose et al., 2025).

Bias is another recurring threat in clinical and assessment settings, such as racial bias in AI-mediated psychiatric diagnosis with large models (Bouguettaya et al.), video interviews (Mujtaba and Mahapatra, 2025), or emotional AI (Chavan et al., 2025).

Deployment discussions increasingly connect to law, clinical workflows, and patient safety. The Draft EU AI Act has implications for profiling (Veale and Zuiderveen, 2021). One must consider the legal risks of AI in mental healthcare (Rahman et al., 2025), as well as safe data management (Raygoza-L et al., 2025). These works collectively push for careful deployment, monitoring, and clinician involvement.

2.1. The DyVaT Algorithm

The DyVaT (Dynamic Variance Thresholding) algorithm is a novel technique designed to reduce the dimensionality of word embeddings in NLP tasks by adaptively removing high-variance noisy dimensions using the cosine distance metric. Using the kneedle algorithm to determine an optimal threshold, DyVaT retains low-variance features that contribute to forming tighter semantic clusters. This process enhances the quality of the vector representations without substantial loss of information, thereby improving downstream tasks such as classification and clustering. Experiments demonstrate DyVaT's ability to generate semantically coherent word collections with less noise compared to other methods, making it a powerful, scalable tool for feature selection in text analysis (Treistman et al., 2022).

3. Methodology

This section presents the project's methodology, including data collection and preprocessing, ML model selection, and system architecture. It outlines the approach used to accurately and reliably classify psychopathic traits from text data.

3.1. Data Collection

The efficacy of the proposed classification framework relies on a structured approach to data handling and feature engineering. To this end, we first define the linguistic corpora used for training and evaluation. Following the dataset description, this section outlines the methodology for generating a semantically dense vocabulary, the subsequent extraction of discriminative features, and the application of synonym-based augmentation to enhance the model's generalizability.

Psychopathic Dataset

- Text samples were gathered from approximately 77 YouTube interviews featuring individuals who were legally and psychologically confirmed as criminal psychopaths.
- The transcripts were generated using the Python library youtube-transcript-api.
- After extraction, a manual review was conducted on all transcripts to verify their accuracy.

Non-Psychopathic Dataset

- The dataset was derived from a validated Kaggle resource, containing transcripts of National Public Radio (NPR) interviews with randomly selected (NPR).
- The non-psychopathic dataset was intentionally selected because it consisted of interview transcripts with a conversational style and structure similar to those in the psychopathic dataset. This deliberate matching ensured consistency in format and context across both datasets, minimizing bias and allowing for reliable comparison of linguistic features.
- The control sample selection was guided by accepted definitions of psychopathy, e.g., "a manipulative, cunning, and antisocial individual who, according to Hare (Hare, 2020), comprises about 1 percent of the general population" (Hancock et al., 2018).
- For this dataset, 10,000 records were initially sampled using two inclusion criteria: (1) each record had to contain a unique EPISODE identifier, and (2) the transcript text was required to include a minimum of 15 sentences. After

applying these criteria, 782 valid records were retained.

3.2. Text Preprocessing

To manage the variability in transcript length, the text was split into uniform segments of fifteen sentences each. Each segment was normalized, cleaned, tokenized, and lemmatized using standard methods (spaCy platform). Short, non-alphabetic or linguistically irrelevant tokens were filtered out (Vasiliev, 2020). Segments with fewer than nine relevant vocabulary terms were discarded to maintain dataset quality and ensure that texts contained sufficient linguistic material for meaningful analysis. Due to an inherent class imbalance in the dataset, data enhancement methods were applied to the minority psychopathic class to equalize representation and prevent biased model learning.

3.3. Vocabulary Construction

The vocabulary construction method (Figure 1) captures key linguistic markers distinguishing psychopathic and non-psychopathic traits. We began with a basic seed vocabulary for each class, manually selecting fifty words per class with the highest coefficients based on our dataset, validated using logistic regression. This vocabulary was then expanded using the DyVaT algorithm. The algorithm increased each class's vocabulary to approximately 1,000 words, and after removing duplicates and applying filtering, the final vocabulary totaled 1,357 words.

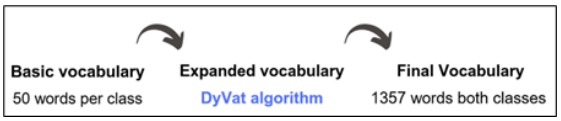


Figure 1: Vocabulary Construction Process

The initial seed vocabulary consisting of fifty terms per class was manually curated. The selection was tuned by both term frequency and the highest logistic regression coefficients observed in our dataset, ensuring that the chosen words captured the most discriminative linguistic markers for each class.

3.4. Expansion via DyVaT Algorithm

DyVaT was applied to expand the seed vocabulary by leveraging semantic similarity from GoogleNews Word2Vec embeddings, using cosine similarity distance thresholding to identify relevant terms. This approach increased the vocabulary coverage to approximately 1,357 words per class while maintaining interpretability. The base vocabulary expansion method relied on a fixed cosine similarity threshold from Word2Vec embeddings, while DyVaT's dynamic variance-based thresholding following initial cosine clustering introduced a broader set of semantically relevant terms, enhancing the model's ability to capture subtle linguistic distinctions. Unlike the base method's static cutoff, which limited coverage and generalization due to its rigid nature, DyVaT effectively balanced expansion with semantic relevance, providing a more robust feature set for psychopathy classification (Treistman et al., 2022).

3.5. DyVaT vs. Base Vocabulary

The visualization compares word expansions generated from the seed word 'control'. In the Base algorithm (Figure 2), related terms are spread with a fixed radius, producing a scattered distribution. In contrast, the DyVaT algorithm (Figure 3) yields tighter clusters of semantically related words, increasing density and better capturing nuanced relationships. This demonstrates that DyVaT provides a more coherent and semantically meaningful expansion compared to the Base method.

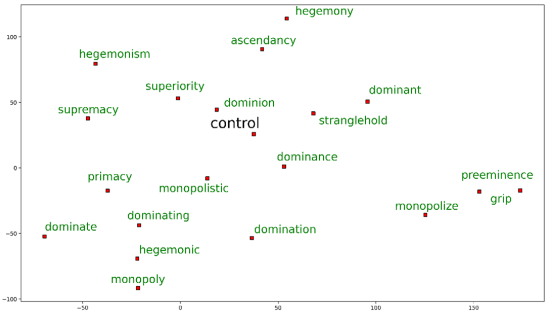


Figure 2: Base Algorithm

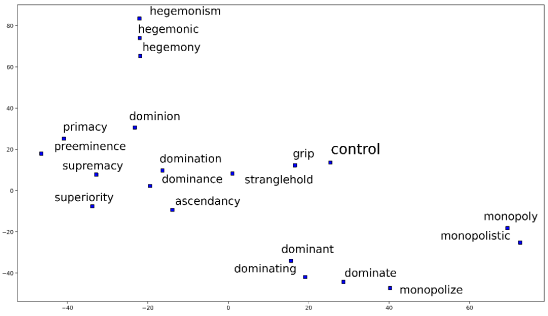


Figure 3: DyVaT Algorithm

3.6. Data Augmentation

To overcome the limitations inherent in small datasets and to improve classification performance, text-based data augmentation techniques were applied to preprocessed text records. Specifically, each original record was transformed, in which selected words, limited only to those present in the

vocab, were replaced by semantically similar synonyms. This was achieved through the use of the SpaCy library, integrating vocabulary vectors and part-of-speech filtering, thus ensuring both contextual fidelity and grammatical correctness. To prevent redundancy, augmented records with high semantic overlap to the originals were systematically removed based on configurable similarity thresholds. This process expanded the dataset from 224 to 436 unique records, resulting in measurable improvements in model performance.

Recent empirical research supports this approach, demonstrating through comprehensive analysis that token-level augmentations, particularly word replacement and random swapping most consistently enhance supervised NLP performance on limited data. These methods generate new text by substituting words or phrases with synonyms from dictionaries or embedding similarities, preserving semantic meaning and original labels while expanding linguistic diversity. Such techniques prove especially effective when training samples are scarce, as they intuitively maintain sentence intent through semantically equivalent token replacements (Chen et al., 2023).

3.7. Feature Extraction

The TF-IDF vectors were restricted to a custom vocabulary, manually curated and further refined using the Dynamic Variance Thresholding (DyVaT) algorithm. Following the DyVaT process, additional manual customization was performed to further optimize the feature set. This ensured that only lexically and semantically meaningful words were included. These vectors served as the input features for machine learning, allowing the algorithm to learn which words and patterns are most indicative of psychopathic versus non-psychopathic text. Feature extraction in this manner provides interpretability and insights into the key linguistic cues distinguishing the two classes.

3.8. Hyperparameter Tuning

Hyperparameter tuning was performed using a grid search approach combined with stratified 5-fold cross-validation to systematically explore combinations of model parameters. The stratification ensured balanced class distributions across folds, addressing potential imbalances in psychopathic versus non-psychopathic samples. The primary optimization objective was the F1-score, chosen to balance precision (reducing false positives) and recall (reducing false negatives), reflecting the critical need to identify all true cases without excessive false alarms.

4. Implementation and Results

Table 1 sumarizes the key differences between the standard baseline and our DyVat augmented approach.

4.1. Baseline vs, Augmented DyVaT Method

Aspect	Baseline	Augmented DyVaT
Pre-processing	Basic noise removal Simple tokenization	Advanced preprocessing including lemmatization Removal of short/irrelevant records
Vocabulary	No extensive feature engineering Relied on basic lexicon if at all	Curated vocabulary built with DyVaT algorithm Expanded manual lexicon to 1,233 semantically relevant features
Feature Extraction	Basic TF-IDF or bag-of-words, no focused filtering	Only records containing =9 VOCAB words retained TF-IDF vectors based on enhanced vocabulary
Data Augmentation	None	Systematic generation of new records by synonym substitution within curated vocabulary
Model Training	Baseline models trained directly on raw data	Models trained and validated on high-quality, filtered, and augmented data
Goals	Quick baselines to establish initial feasibility	Maximized performance, interpretability, and generalization (based on literature insights and pilot results)

Table 1: Comparison Between Original and Final Methods

4.2. Original vs. Augmented Dataset

Original Dataset Contained raw transcribed interviews labeled according to psychopathic and non-psychopathic traits, without advanced filtering or augmentation.

Augmented Dataset Expanded version created by applying the full augmentation pipeline, resulting in increased linguistic diversity, a more

balanced class distribution, and improved robustness for model training and evaluation.

4.3. Data Processing Pipeline

The data processing pipeline was systematically developed to ensure that raw interview data was transformed into high-quality, interpretable features optimized for ML classification. The pipeline includes these key stages:

Data Ingestion and Organization

Raw interview transcripts were extracted directly from MongoDB and imported into pandas DataFrames, establishing a structured, accessible format for all downstream analyses.

Text Preprocessing

- The transcript texts underwent thorough normalization and noise reduction, including lowercasing, removal of punctuation, stop words, and irrelevant symbols.
- Standardized input size by dividing transcripts into fixed-length segments.
- Further linguistic cleaning was completed by tokenization and lemmatization (via spaCy), transforming words to their canonical forms and ensuring linguistic consistency.
- Semantic filtering excluded segments with insufficient relevant vocabulary, minimizing noise and ensuring only linguistically meaningful samples advanced.

Vocabulary Construction And Restriction

Feature extraction was restricted to a carefully engineered vocabulary, expanded using the DyVaT algorithm. This method augmented a manual seed list with semantically similar terms, enhancing lexical coverage for both psychopathic and non-psychopathic classes while ensuring feature interpretability.

Data Augmentation And Balancing

Data augmentation strategies were employed for the minority class, generating new records by substituting words with synonyms within segments, thus increasing dataset size and diversity while preserving semantic integrity.

Feature Extraction

Texts were vectorized using the TF-IDF algorithm, resulting in numerical features that represent each record in terms of word importance and discriminative value.

Model Training, Selection and Evaluation

- Multiple ML models were trained using stratified k-fold cross-validation for robust and balanced evaluation.
- Model persistence was achieved by serializing the optimal models with joblib, supporting reproducibility and deployment.

Feature Importance And Visualization

- The key linguistic features distinguishing psychopathic vs. non-psychopathic speech were identified through LR coefficient analysis.

- The fifty most influential words per class were visualized as word clouds, highlighting dominant lexical markers and supporting transparent model interpretation.

4.4. Feature Importance

The analysis of the most influential features was conducted using the coefficient values derived from a logistic regression model, which allowed for clear interpretability of the linguistic markers associated with psychopathy. The logistic regression coefficients quantify the contribution of each feature to the classification decision by indicating the change in the log-odds of the target class.

Figures 4 and 5 present the top 50 linguistic features most strongly associated with each class (Psychopath vs. Non-Psychopath).

Figure 4: Top 50 Features of Psychopaths

Figure 5: Top 50 Features of Non-Psychopaths

5. Analysis

5.1. Baseline Method

The baseline model (Table 2) was trained without dataset balancing, without the DyVaT vocabulary, and without any filtering or augmentation. The TF-IDF vectorizer relied entirely on the raw, unfiltered lexical space, resulting in high noise levels and severe class imbalance.

This configuration struggled to capture meaningful psychopathy-related linguistic patterns, highlighting the need for improved data quality and class balance.

Model	F1-score	Vocabulary Size
SVM	0.5000	2881
Logistic Regression	0.3871	80182
Random Forest	0.3200	43028

Table 2: Performance Metrics for Baseline Method

- Training set size: 781
- Test set size: 194
- Total records: 975
- Psychopath records: 193
- Non-psychopath records: 782

5.2. Augmented DyVaT Method

Our method uses data augmentation was combined with DyVaT expansion (Table 3 and Figure 6), expanding the psychopathic class and enhancing generalization. Linear SVM achieved the highest performance. This result represents the culmination of progressive refinement, demonstrating that the combination of controlled vocabulary, balanced classes, semantic filtering, and targeted augmentation yields the most reliable psychopathy detection model.

Model	F1-score	Vocabulary Size
SVM	0.7957	1233
Logistic Regression	0.6882	1233
Random Forest	0.7391	1233

Table 3: Performance Metrics for Augmented DyVaT Model

- Training set size: 407
- Test set size: 99
- Total records: 506
- Psychopath records: 253
- Non-psychopath records: 253

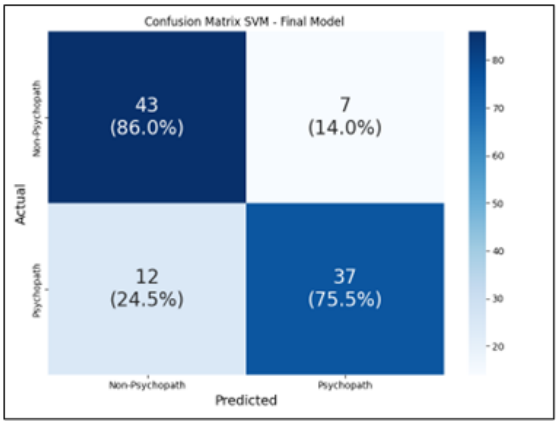


Figure 6: Confusion Matrix of the Augmented DyVaT Method

6. Discussion

Our research successfully achieved its objectives by developing a ML and NLP-based system capable of identifying psychopathic traits through linguistic analysis. After systematically evaluating multiple methods, the Support Vector Machine (SVM) trained with the DyVaT vocabulary and augmented dataset was identified as our optimal algorithm for psychopathy detection.

This model achieved the highest F1-score (0.7957), while maintaining strong generalization and interpretability. The integration of longer text records, controlled vocabulary filtering, and targeted data augmentation proved essential for capturing psychopathy-related linguistic structures without overfitting. The confusion matrix further demonstrated that the SVM achieved a high true positive rate while maintaining a low false positive rate, reflecting strong discriminative ability and balanced performance.

Thus, the findings confirm that computational methods, particularly SVM with carefully constructed linguistic features, are effective for early detection of psychopathic traits. These results strengthen the positive consideration of ML approaches in psychological assessment.

Future work may focus on expanding the dataset, incorporating deep learning architectures, integrating sentiment and emotion analysis, and exploring multimodal inputs (e.g., audio, video) to further improve detection accuracy and robustness.

6.1. Project Challenges and Limitations

During the execution of this research, several major challenges were identified, which influenced the planning, implementation, and evaluation stages:

Ambiguity in Defining Psychopathy

There is no universally accepted definition of psychopathy. Diagnosis relies on a complex combination of psychological, behavioral, and social criteria. Although academic literature offers several assessment tools, such as the Psychopathy Checklist-Revised (PCL-R), there is no full consensus on the construct's boundaries. This lack of uniformity complicates the process of labeling training data and may negatively impact model performance.

Dataset Collection and Preparation

Censored sources: Some available materials, such as YouTube recordings or interview transcripts, included censorship, omissions, or edits that reduced data completeness and reliability.

Scarcity of psychopathy-labeled texts: Texts authored or spoken by clinically confirmed psychopaths are rare, due to ethical, privacy, and data access constraints.

Non-psychopath text collection: This task presented additional complexity, as sufficiently long texts authored or spoken by non-psychopaths were difficult to obtain. Furthermore, care had to be taken to ensure that the selected individuals represented a general population rather than a narrowly defined or homogeneous group, in order to reduce potential sampling biases.

Selecting And Evaluating ML Model

Developing an accurate and reliable psychopathy detection model posed significant challenges, particularly in the initial stages of algorithm selection. It was unclear whether to begin with traditional ML algorithms or to employ deep learning techniques. Consequently, careful experimentation and systematic evaluation were required to determine the optimal modeling strategy, balancing predictive performance with computational efficiency and training feasibility.

Building A Contexual Vocabulary

Another challenge in this study was determining whether to use a predefined vocabulary consisting of words specifically associated with psychopathic and non-psychopathic speech, or to derive the vocabulary directly from the training dataset. Careful consideration was required to determine how to construct this vocabulary in a manner that maximizes its discriminative power, captures relevant linguistic patterns for each class, and avoids introducing bias or overfitting.

6.2. Commercial and Societal Value

The developed system has potential commercial applications in security, mental health, and criminology, providing tools for early risk assessment and decision support. Beyond commercial value, the research contributes to societal well-being by enabling more proactive identification of high-risk individuals and supporting preventive interventions.

6.3. Summary

This research set out to determine whether psychopathic linguistic patterns can be identified through NLP and machine learning, with the goal of creating a scalable, interpretable, and non-invasive detection framework. By constructing a novel dataset from verified psychopathic and non-psychopathic interviews, expanding vocabulary coverage using the DyVaT algorithm, and applying targeted preprocessing, semantic filtering, and data augmentation, the study achieved robust classification results.

The Linear SVM model emerged as the optimal solution, delivering an accuracy of 0.8031 and an

F1 score of 0.7957, outperforming alternative models such as Logistic Regression and Random Forest. These results underscore that psychopathy-related language patterns are consistent and detectable when captured through a well-engineered lexical feature space. Moreover, the interpretability of the Linear SVM provides valuable transparency-an essential quality in forensic, legal, and clinical applications.

The contributions of this study are threefold:

- **Data Contribution** Development of a real-world, verified dataset of psychopathic and non-psychopathic speech, providing a foundation for future research.
- **Methodological Innovation** Application of DyVaT for vocabulary expansion, enhancing feature quality while maintaining explainability.
- **Practical Relevance** Validation of an NLP-based detection tool that could support early risk assessment in mental health, security, and law enforcement contexts.

While the findings are promising, they should be interpreted in light of the study's limitations, including dataset size, English-only scope, and reliance on text transcripts without paralinguistic data. These constraints point toward several avenues for future research: expanding the dataset, incorporating multilingual sources, integrating multimodal features, and leveraging advanced deep learning architectures such as transformer-based models.

In essence, this project demonstrates that language serves not only as a medium for communication but also as a measurable indicator of underlying personality traits. Through careful design and rigorous validation, the framework developed here shows the potential of computational psycholinguistics to aid in the early identification of psychopathy-supporting, rather than replacing, professional judgment in high-stakes environments.

7. References

Jonathan Adkins, Ali Al Bataineh, and Anthos Khanal. 2025. A psycholinguistic nlp framework for forensic text analysis of deception and emotion. *Frontiers in Artificial Intelligence*, 8:1669542.

Saqib Alam and Nianmin Yao. 2019. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, 25(3):319–335.

Hesham Allam, Lisa Makubvure, Benjamin Gyamfi, Kwadwo Nyarko Graham, and Kehinde Akinwolere. 2025. Text classification: How machine learning is revolutionizing text categorization. *Information*, 16(2):130.

Yehia Ibrahim Alzoubi, Ahmet E Topcu, and Ahmed Enis Erkaya. 2023. Machine learning-based text classification comparison: Turkish language context. *Applied Sciences*, 13(16):9428.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, Gregory Webster, and Damon Woodard. 2025. Psytex: A knowledge-guided approach to refining text for psychological analysis. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 151–178.

Evan Bose, Chaitanya Anil Kumar, and N Meenakshi. 2025. Ai-driven psychological profiling on social media: Mechanisms, ethical breaches, and regulatory challenges in data inference. recent trends in social studies. 2025; 2 (1): 1–7p. *AI-Driven Psychological Profiling on Social Media Bose et al. STM Journals*, page 2.

A Bouguettaya, EM Stuart, and E Aboujaoude. Racial bias in ai-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. npj digit. med. 8, 332 (2025).

Alexis Carrillo, Simon Friedrich Roske, Rebeca Ianov-Vitanov, Enrico Perinelli, Alessandro Grecucci, and Massimo Stella. 2025. Textual forma mentis networks bridge language structure, emotional content and psychopathology levels in adolescents. *arXiv preprint arXiv:2505.06387*.

Vivek Chavan, Arsen Cenaj, Shuyuan Shen, Ariane Bar, Srishti Binwani, Tommaso Del Becaro, Marius Funk, Lynn Greschner, Roberto Hung, Stina Klein, et al. 2025. Feeling machines: Ethics, culture, and the rise of emotional ai. *arXiv preprint arXiv:2506.12437*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Mamata Das, PJA Alphonse, et al. 2023. A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset. *arXiv preprint arXiv:2308.04037*.

Krishna Kant Dixit, Sumit Pundir, Anurag Shrivastava, C Praveen Kumar, Arun Pratap Srivastava, and Pankaj Singh. 2023. Analyzing textual data for mental health assessment: Natural language processing for depression and anxiety. In *2023 10th IEEE Uttar Pradesh Section International*

Conference on Electrical, Electronics and Computer Engineering (UPCON), volume 10, pages 1796–1802. IEEE.

Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. 2022. A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83.

Barbara Gawda. 2022. The differentiation of narrative styles in individuals with high psychopathic deviate. *Journal of Psycholinguistic Research*, 51(1):75–92.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2025. Improving suicidal ideation detection in social media posts: Topic modeling and synthetic data augmentation approach. *JMIR Formative Research*, 9:e63272.

Yuting Guo, Anthony Ovadje, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10):2181–2189.

J. T. Hancock, M. Woodworth, and R. Boochever. 2018. Psychopaths online: The linguistic traces of psychopathy in email, text messaging and facebook. *Media and Communication*, 6(3):83–92.

R. D. Hare. 2020. The pcl-r assessment of psychopathy. In *The Wiley International Handbook on Psychopathic Disorders and the Law*, pages 63–106. Wiley.

Jashanjot Kaur and P Kaur Buttar. 2018. A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4):207–210.

Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. 2022. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*, 10.

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. Evaluating psychological safety of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1826–1843.

Jingfang Liu, Peng Ding, and Jie Chen. 2025. Depglm: Depression degree recognition on social media based on large language models. *Digital Health*, 11:20552076251408281.

Huimin Mao and Qing Han. 2025. Enhancing textgcn for depression detection on social media with emotion representation. *Frontiers in Psychology*, 16:1612769.

Leberecht Maxim, Nedderhoff Andre, Zitzmann Steffen, and Hecht Martin. 2025. Comparing machine learning methods for predicting dark triad personality traits using social media text data. *Journal of Research in Personality*, page 104690.

Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507.

Dena F Mujtaba and Nihar R Mahapatra. 2025. Behind the screens: Uncovering bias in ai-driven video interview assessments using counterfactuals. *arXiv preprint arXiv:2505.12114*.

Anam Naz, Hikmat Ullah Khan, Amal Bukhari, Bader Alshemaimri, Ali Daud, and Muhammad Ramzan. 2025. Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. *Artificial Intelligence Review*, 58(8):239.

NPR. Npr media dialog transcripts. Kaggle Dataset. Accessed: September 22, 2025.

Dhruv Patel and Anju Johnson. 2025. Detecting narcissistic personality disorder (npd): A hybrid regex and nlp based ai approach with phase-aware classification. *IEEE Access*.

Rilo Chandra Pradana and Derwin Suhartono. 2024. Synonym replacement augmentation for handling data imbalance in personality classification. In *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 1–6. IEEE.

Md Ashrafur Rahman, Evangelos Victoros, Rob Davis, Tariq Duaa, Yeasna Shanjana, and Md Rabiul Islam. 2025. Use of artificial intelligence in mental healthcare, health psychology, and related research: A narrative review to address challenges and opportunities. *Health Science Reports*, 8(12):e71595.

María E Raygoza-L, Jesús Heriberto Orduño-Osuna, Roxana Jimenez-Sanchez, and Fabian N Murrieta-Rico. 2025. Innovative artificial intelligence approaches for identifying and managing dsm cluster b personality disorders in mental health: A case study on the dark triad. In *Exploring Psychology, Social Innovation and Advanced Applications of Machine Learning*, pages 1–20. IGI Global Scientific Publishing.

Minhah Saleem and Jihie Kim. 2024. Intent aware data augmentation by leveraging generative ai

for stress detection in social media texts. *PeerJ Computer Science*, 10:e2156.

Bosubabu Sambana, Kondreddygari Archana, Suram Indhra Sena Reddy, Shaik Meethaigar Jameer Basha, and Shaik Karishma. 2025. Data augmentation for cognitive behavioral therapy: Leveraging ernie language models using artificial intelligence. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 204–209. IEEE.

Lingzhi Shen, Yunfei Long, Xiaohao Cai, Guanming Chen, Yuhan Wang, Imran Razzak, and Shoaib Jameel. 2025. Ll4g: Self-supervised dynamic optimization for graph-based personality detection. *arXiv preprint arXiv:2504.02146*.

Zhivar Sourati, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Nuan Wen, Ala Tak, Fred Morstatter, and Morteza Dehghani. 2024. Secret keepers: The impact of llms on linguistic markers of personal traits. *arXiv preprint arXiv:2404.00267*.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. 2025. Challenging the validity of personality tests for large language models. In *Proceedings of the 2025 Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 74–81.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3117–3133.

Avraham Treistman, Dror Mughaz, Ariel Stulman, and Amit Dvir. 2022. Word embedding dimensionality reduction using dynamic variance thresholding (DyVaT). *Expert Systems with Applications*, 208:118157.

Yuli Vasiliev. 2020. *Natural Language Processing with Python and spaCy*. Packt Publishing. Accessed: December 6, 2025.

Michael Veale and Borgesius Frederik Zuiderveen. 2021. Demystifying the draft eu artificial intelligence act. *Computer Law Review International*, 22(4):97–112.

G Venkateshwarlu, S Akhila, V Kavyasree, S Vishnu, and VS Prasad. 2024. Enhanced text classification using random forest: Comparative analysis and insights on performance and efficiency. *Int. J. Comput. Eng. Res. Trends*, 11:1–8.

Sumona Yeasmin, Nazia Nowshin, and Tasnia Afrin Chowdhury. 2024. Identifying human dark triad from text data through machine learning models. *International Journal of Research and Innovation in Applied Science*, 9(6):89–104.

Jianlong Zhou and Fang Chen. 2023. Ai ethics: From principles to practice. *Ai & Society*, 38(6):2693–2703.

Developing Annotation Guidelines for CSAM Prevention Interventions: Psychosocial Risk and Protective Factors Grounded in Research and Clinical Practice

Vera Czehmann[1,3], Christine Hovhannisyan[4], Lena Hoffmann[3], Paula Busch[2], Ibrahim Baroud[1,3], Sebastian Möller[1,3], Roland Roller[1], Hannes Gieseler[2], Lisa Raithel[1,3,5,6]

[1]German Research Center for Artificial Intelligence (DFKI GmbH),
[2]Charité – Universitätsmedizin Berlin, Institute of Sexology and Sexual Medicine,
[3]Quality & Usability Lab, Technische Universität Berlin, [4]Humboldt-Universität zu Berlin,
[5]BIFOLD – Berlin Institute for the Foundations of Learning and Data,
[6]Charité – Institut für Künstliche Intelligenz in der Medizin (IKIM)
Berlin, Germany
{vera.czehmann, roland.roller}@dfki.de
christine.hovhannisyan@student.hu-berlin.de
l.hoffmann@campus.tu-berlin.de
{paula.busch, hannes.gieseler}@charite.de
{ibrahim.baroud, raithel}@tu-berlin.de

Abstract

This work discusses sexual offending, specifically child sexual abuse material (CSAM), in the context of prevention. We introduce a domain-specific, span-level annotation scheme and guidelines to identify psychosocial *risk* and *protective* factors in therapist-led, anonymous chat interventions with voluntarily help-seeking individuals concerned about their pedophilic interests and the risk of CSAM use. The scheme is grounded in previous research and clinical experience, and intended for within-intervention guidance and longitudinal tracking, rather than actuarial risk scoring. Annotating a pilot subset (8 clients, 31 sessions), inter-annotator agreement was moderate but improved after calibration, which is consistent with the linguistic and clinical ambivalence present in the data. We track a session-wise *Protective Ratio*, i.e., the share of protective factors among all coded factors, and examine its behaviour over time during the intervention and around self-reported relapse within clients. In exploratory automation, LLM-based span extraction outperforms BERT baselines but overall performance remains limited by small data and mixed-evidence spans. While complete anonymisation of the corpus is in progress, we release the label scheme, guidelines, and non-sensitive artefacts of our analyses.

Keywords: CSAM prevention, psychosocial risk and protective factors, annotation guidelines, span extraction, grounded methodology, therapist expertise

1. Introduction and Motivation

Sexual interest in minors (commonly described as pedophilic or hebephilic interest) constitutes a persistent sexual preference pattern rather than, in itself, a criminal act (Jahnke, 2018). Such interests are classified under paraphilic disorders only when the person has acted on their urges, they cause distress, impairment, or involve risk of harm to themselves or others (American Psychiatric Association, 2022). Acting on such urges can manifest in different forms, including contact child sexual abuse (CSA) and the use of child sexual abuse material (CSAM); images or other media depicting the abuse or exploitation of minors. While these behaviours are interconnected, research recognises distinctions between contact and non-contact offending populations in terms of risk profiles, motivations, and intervention needs (Babchishin et al., 2015).

A growing body of research also distinguishes between forensic populations, who enter the system following detected CSAM-related offences, and clinical populations. Within this so-called *Dunkelfeld* (the "dark figure" of undetected offences), there exists a subset of individuals actively seeking help to prevent CSAM-related (re)offending (Von Franqué et al., 2023). In our clinical context, we refer to this as *relapse*, i.e., a return to CSAM use, as self-reported by clients. Many such individuals experience fear of disclosure (Jahnke, 2018). In response, several prevention-oriented programmes have emerged to offer confidential and, in some cases in their online extensions, anonymous therapeutic or self-guided interventions for individuals seeking to manage their attraction responsibly. These initiatives reveal an underserved population of help-seeking individuals whose communications with counsellors provide a unique window into cognitive, emotional and behavioural cues relevant for relapse prevention. Understanding which factors could predict positive or negative trajectories is critical for effective intervention.

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 126–145
12 May 2026.

We differentiate between risk factors that could increase the likelihood of relapse, and protective factors, resources, strategies, or cognitions that could help prevent or reduce the likelihood of relapse. This work explores how such factors are expressed linguistically in preventive CSAM intervention chats and whether they can be tied to intervention outcome and self-reported relapse behaviours. We introduce a domain-specific annotation scheme tailored to these dialogues and report inter-annotator agreement and statistical analyses on detected factors.

The scheme is intended to support the manual creation of a gold standard set that could train and evaluate automatic classifiers to produce a large-scale, transparent corpus. Prospectively, this data could be utilised in building a tool to help flag risk and protective factors in client talk, support therapists' decision-making during sessions and enable the longitudinal tracking of client progress with respect to therapist intervention. We furthermore provide a *first baseline using Large Language Models for extracting the defined factors*. Finally, we conducted a *semi-structured expert interview*, as previously done by Klymenko et al. (2022). References to the interview or expert opinion will be marked throughout with an asterisk (*). The full annotation scheme and guidelines and a transcript of the interview can be found on Zenodo[1].

2. Related Work

Unlike other medical domains with routinised outcomes and large public datasets, preventive therapeutic interventions regarding sexual offences lack shareable corpora, despite increasing interest in dynamic risk modelling and online interventions. Therefore, this work bridges three strands of prior work: (i) client language and outcomes, (ii) forensic risk-assessment constructs, and (iii) secondary prevention in the *Dunkelfeld*.

Client language and intervention outcomes. A recent survey of mental health datasets emphasises that while labelled datasets on multi-class classification and questionnaire score prediction exist, there is a particular scarcity in genuine therapy corpora (Mandal et al., 2025). Research on online, text-based counselling shows that client language can be annotated reliably and is predictive of subsequent outcomes. For Motivational Interviewing (MI), a therapy strategy used, e.g., in the treatment of addiction, Wu et al. (2023) released a dataset of transcribed counselling dialogue demonstrations, expert-annotated for MI-specific concepts such as change- and sustain-centered client talk on the dialogue and utterance levels. Previous works in the field of MI have found that while an association with change talk has not been consistently reported across studies, sustain talk was positively associated with worse outcome (Magill et al., 2018).

Ewbank et al. (2021) manually coded transcripts of internet-enabled Cognitive Behavioural Therapy (CBT) for five categories of client utterances, informed by the MI technique (Amrhein et al., 2003). This was then utilised in training a deep-learning classifier to auto-code transcripts at scale. Model performance reached human-level agreement on most of the categories and, crucially, the automatically derived signals were reliably linked to outcomes. They also identified demographic predictors of reliable improvement from the first session. Together, these findings indicate that span-level labels on client talk are both feasible and informative for downstream support tools.

Risk assessment in the forensic field. Forensic risk-assessment frameworks used with convicted sexual offenders offer constructs relevant to, among others, sexual preoccupation, cognitive distortions, and self-regulation. Actuarial tools (e.g., STATIC-99R) quantify risk from static, file-based information (Phenix et al., 2017). Structured professional judgement (SPJ) approaches (SVR-20; RSVP) use standardised item sets in combination with decision guidance to support clinician risk formulation (Hart and Boer, 2020). Dynamic instruments (STABLE-2007; ACUTE-2007) capture relatively stable and acute changeable factors, and recent work links their scores to recidivism among men adjudicated for CSAM offences (Babchishin et al., 2023). For CSAM-specific recidivism, the Child Pornography Offender Risk Tool (CPORT) operationalises offence-relevant items and shows predictive utility (Seto and Eke, 2015). Recent review work has also synthesised psychosocial characteristics and risk-related profiles in detected CSAM offenders (Barroso et al., 2026). These tools and findings conceptually inform our annotation scheme. However, they are not directly transferable to anonymous, therapist-led prevention chats with undetected offenders or individuals at risk of CSAM-related offending (Von Franqué et al., 2023).

Secondary prevention in the *Dunkelfeld*. Secondary prevention services reach voluntarily help-seeking individuals outside the justice system. Evaluations report decreases in offence-supporting cognitions and emotional deficits, alongside gains in self-regulation, although a residual risk of CSAM use (relapse) may persist (Beier et al., 2015; Von Franqué et al., 2023). Therapist-led, chat-based anonymous interventions broaden access, with CBT approaches showing initial reductions in

[1] https://zenodo.org/records/19189153

CSAM viewing among motivated users (Lätth et al., 2022). Other studies indicate substantial demand and characterise user profiles, including factors linked to help-seeking and motivation to stop (Insoll et al., 2024), as well as higher distress and CSAM use disclosures among users with pedophilic or hebephilic interests (Schuler et al., 2021).

Prototype therapist-assistive AI tools for chat-based interventions in the domain reduce clinicians' perceived cognitive load (Deshpande et al., 2025a), and retrieval-augmented LLM suggestions can match or exceed therapist replies (Deshpande et al., 2025b). These findings underscore that therapist support at scale requires machine-actionable, span-level labels that index risk or protective factors.

3. Data

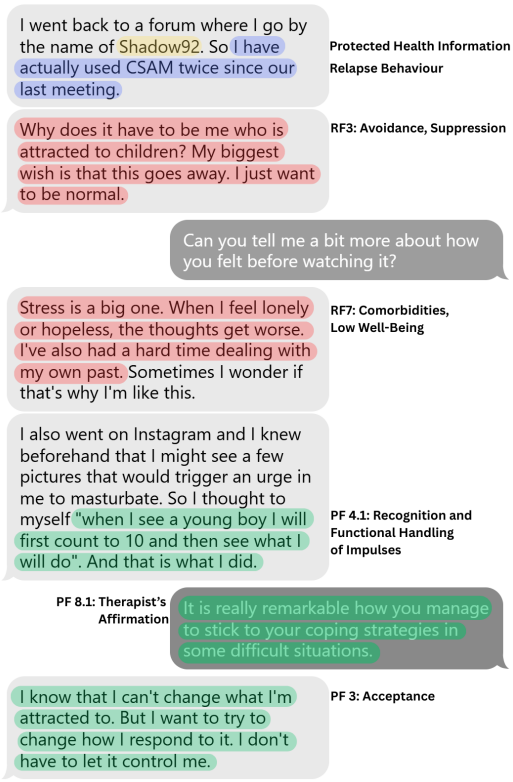


Figure 1: A fictional example excerpt of a session between a therapist and a client with annotated risk and protective factors.

We tailored our annotation scheme to chat transcripts from intervention chats between help-seeking individuals and therapists. They were collected within an anonymous online study on preventive support for individuals self-reported to be at risk of CSAM-related offending, conducted by the Institute of Sexology and Sexual Medicine at Charité – Universitätsmedizin Berlin. A fictional example of one such therapy chat is shown in Figure 1.

Clients	Sessions	Tokens	Tokens per Session
8	31	62,034	2,001

Table 1: Corpus size and typical session length.

Participants enrol voluntarily and have the right to withdraw at any time. Eligible individuals provide informed consent to the use of their chat transcripts for research. Over the course of a 12 week period, participants receive access to self-help material and a varying amount of scheduled chat sessions with a therapist of 50 minutes each. Both therapists and clients remain anonymous throughout. Chat logs are stored on secure institutional servers without personal identifiers and are automatically de-identified, replacing detected PHI with placeholders before using them for annotation. We retain anonymised speaker and time metadata to capture turn-taking.

Although a substantially larger, multilingual corpus is being collected, for this first annotation study we focus on a randomly selected subset of eight English speaking clients (details in Table 1). While we cannot release raw transcripts at this time, we are conducting iterative de-identification to annotate and remove remaining Protected Health Information[2] as well as indirect identifiers (as suggested by Baroud et al., 2025) with the goal of sharing an extensively anonymised subset in the future, subject to ethics approval and data sharing agreements.

4. Guideline Development

We next describe the process of developing our annotation scheme and guidelines. Factors are intended as clinically interpretable anchors and actionable labels for session guidance and longitudinal tracking, not as actuarial or forensic risk scores.

4.1. Annotation Scheme and Guidelines

We target a preventive, non-forensic setting: voluntarily help-seeking individuals concerned about their (risk of) CSAM use. Our proposed annotation scheme is shown in Table 2. The first version was drafted by a team member with extensive clinical experience working with individuals at risk of CSAM-related offending in both offline and online settings. Combined with this, a targeted synthesis of prior research informed a first-pass taxonomy of linguistically observable risk and protective factors indicative of potential relapse behaviour, suitable for span-level annotation in therapist-led chats. In our annotation, we prioritised concepts that clinicians actively monitor during prevention work, and which can be expressed explicitly or implicitly in

[2]PHI, annotation guidelines by Lohr et al. (2024).

client talk and could be actionable for session-by-session tracking.

The complete annotation guidelines include further descriptions of factors, fictional example spans for each factor, and guides on prioritisation between some semantically adjacent factors. Definitions of some factors also include client talk of recognising the importance of certain protective concepts, even if it did not reflect in their actions. While the proposed annotation scheme is domain-specific and purpose-built for preventive chat interventions, all factors are conceptually grounded in prior research.

4.2. Grounding in Existing Literature

Sexual preference for children (RF1) has been found to be more prevalent in users of CSAM, with previous work reporting that online offenders score higher on measures of pedophilic interest (Schuler et al., 2021; Babchishin et al., 2015). Regarding **preference patterns** (RF1.1), the CPORT, having shown predictive utility for sexual recidivism among CSAM offenders, includes items on male-focused interest as a risk-enhancing factor (Seto and Eke, 2015). Consequently, **non-exclusivity**, present sexual interest in adults (PF1), is treated as a protective factor consistent with strengths-based rehabilitation (Ward et al., 2025; Willis and Ward, 2024) and emerging work on dynamic protective factors that reduce reliance on illegal material by opening pathways to viable, lawful intimacy (Thornton et al., 2024).

The Good Lives Model (Ward et al., 2025) frames capability-building for prosocial, consensual relationships as incompatible with offending. Thus, mentions of **healthy intimacy** (PF2) mark a protective trajectory. Clinic-centred prevention reports similarly target relational functioning and empathy as change mechanisms in voluntarily help-seeking populations (Von Franqué et al., 2023; Beier et al., 2015). **Difficulties forming or maintaining trusting, mutually supportive adult relationships** (RF2, 2.1) are treated in structured professional judgement (SPJ) and dynamic frameworks as relatively stable vulnerabilities (Babchishin et al., 2023; Hart and Boer, 2020).

Non-acceptance, active avoidance or suppression of the sexual preference in client statements (RF3) is treated as risk relevant, as it can undermine self-regulation and increase distress, making planned coping harder (Hart and Boer, 2020). Shame- and avoidance-driven presentations could also reduce engagement and follow-through, with clinical prevention reports suggesting that moving towards an **accepting, integrated self-image** (PF3) can support consistent coping and value-congruent behaviour (Von Franqué et al., 2023; Beier et al., 2015).

SPJ frameworks consider problems with stress or coping a risk factor in the psychological adjustment domain (Hart and Boer, 2020). Consistent with this, we annotate **dysfunctional coping and impulse control deficits** (RF4, 4.1), because relapses often occur when self-regulation fails in the presence of acute triggers*. CBT models describe the operational mechanism of high-risk situations paired with ineffective coping leading to relapse, whereas specific, **healthy coping** responses (PF4) can interrupt the chain (Marlatt and Donovan, 2005). In the CSAM context, early evidence from internet-delivered CBT shows that equipping motivated users with concrete coping strategies can reduce viewing of problematic content (Lätth et al., 2022).

Cognitive validation (RF5) captures offence-supportive attitudes and minimisations that reduce perceived wrongfulness or harm, which has been found to be an empirically supported risk factor for sexual recidivism (Mann et al., 2010). SPJ frameworks also consider denial of sexual violence and attitudes that condone violence as risk factors (Hart and Boer, 2020). Qualitative interviews suggest that there are implicit theories in child abusers that account for the majority of their cognitive distortions (Marziano et al., 2006). In CSAM specific work, reviews report closely related cognitions in online offending (Bartels and Merdian, 2016). We code **recognition of harm** (PF5) as protective counterpart.

Past problematic behaviour, specifically CSAM use, and **criminal history** (RF6) are consistently linked to higher recidivism risk in CSAM literature (Babchishin et al., 2015). The CPORT operationalises case file-derived predictors (e.g., offence history) and has demonstrated predictive utility for sexual recidivism among CSAM offenders (Helmus et al., 2025; Eke et al., 2019). We code as protective factors (PF6, 6.1) when clients articulate **healthy sexual behaviour** or **abstinence from illegal material explicitly for fear of legal consequences**, reflecting observations that such client talk may relate to more favourable trajectories*.

Comorbidities and low well-being (RF7) are annotated because co-occurring burdens (e.g., depressed mood, anxiety, or substance use) can weaken self-regulation and amplify triggers, increasing relapse risk*. This is consistent with instruments that flag negative affect and substance use (Babchishin et al., 2023) and with SPJ guidance integrating psychosocial adjustment and mental disorders into risk assessment (Hart and Boer, 2020). Complementarily, we annotate explicit statements of **well-being** (PF7) that go beyond courtesy phrases.

Poor therapy and change commitment, and externalisation (RF8, 8.1) captures general agency-distancing client talk. SPJ guidance treats

Risk Factors	
RF0	Not specified
RF1	Sexual preference for children
RF1.1	Sexual preference for male children
RF2	Intimacy deficits, lack of trustful social relationships
RF2.1	Being in a dysfunctional relationship
RF3	Avoidance, suppression, missing acceptance of the sexual preference
RF4	Dysfunctional coping (strategies)
RF4.1	Lack of impulse control
RF5	Reduction of cognitive dissonance, cognitive validation of problematic behaviour
RF6	Past problematic behaviour, criminal history
RF7	Comorbidities, low well-being, other psychological problems or diseases
RF8	Poor therapy and change commitment, missing confidence about reaching the goals and changing behaviour
RF8.1	Directing responsibility to someone else, externalising problems
RF9	Sociodemographic factors
RF10	Hypersexuality, sexual preoccupation
RF10.1	Failure to satisfy sexual needs in a healthy way
RF11	Hostility, preoccupation towards other groups
RF12	Compulsive sexualisation of non-sexual content (of children) and/or situations (with children)

Protective Factors	
PF1	Non-exclusivity of the sexual preference for children
PF2	Healthy intimacy, trustful social relationships and acknowledgement of the importance of it
PF3	Acceptance of the sexual preference
PF4	Functional coping (strategies)
PF4.1	Recognition and functional handling of impulses
PF5	Recognition of the abuse of children in the material
PF6	Healthy sexual behaviour
PF6.1	Abstinence of problematic behaviours due to fear of legal consequences
PF7	Well-being
PF8	Cooperation with therapist, commitment to the treatment or study setting
PF8.1	Therapist's affirmation of commitment
PF10	Skills to satisfy sexual needs (urges) in a healthy way without harming themselves and/or others

Table 2: Overview of the proposed annotation scheme with risk (left) and protective (right) factors.

issues with treatment or supervision and negative attitudes towards the intervention as risk indicators (Hart and Boer, 2020). In MI research, sustain talk and therapist-client discord track poorer outcomes, whereas **commitment** (PF8) and reason or need language predict improvement (Miller, 2023; Magill et al., 2018). Analyses on internet-enabled CBT similarly show that motivated client talk relates to better outcomes (Ewbank et al., 2021). Specific to chat-delivered intervention, we also code **therapist's affirmation of commitment** (PF8.1).

We consider several risk-relevant **sociodemographic factors** (RF9). While the CPORT considers age of 35 or younger at time of the index investigation as an item related to higher risk of recidivism (Seto and Eke, 2015), we follow the observations of involved therapists that relatively young age at the time of the intervention could be associated with negative outcomes*. Informed by SPJ guidance, we also annotate problems with employment (Hart and Boer, 2020), and housing.

In risk factor syntheses, **hypersexuality** (RF10) is identified as a correlate (Mann et al., 2010), and CSAM-focused reviews note how sexual preoccupation, in combination with availability and anonymity online, sustains use and can co-occur with escalation of the unwanted behaviour in severity or frequency (RF10.1) (Helmus et al., 2025; Baskurt et al., 2025). Correspondingly, we explicitly code mentions of **skills to satisfy sexual needs in a healthy way** without harming themselves or others (PF10), countering preoccupation by enabling concrete, value-consistent choices (Lätth et al., 2022).

Hostility-laden preoccupation (RF11), the combination of hostile affect and sexual focus, is theoretically well-grounded. The confluence model links hostile masculinity and impersonal or **sexualised cognition** to elevated risk for sexual aggression (Malamuth et al., 1996). We code this as compulsive sexualisation of non-sexual content or situations (RF12).

5. Annotation

The annotation was carried out by a team of five trained annotators. The team included both domain experts and trained research assistants, ensuring that clinical expertise and methodological

rigor were combined throughout the annotation process. One of the annotators was a physician and experienced therapist actively involved in the chat interventions from which the data were drawn, providing first-hand contextual insight into the therapeutic setting. The remaining annotators comprised a research associate in clinical psychology, a final year Bachelor's student in psychology, a final year Master's student in computer science, and a computer science student with a medical background.

All annotators were fluent in English and received detailed instruction on the annotation guidelines prior to beginning the task. Before the main annotation phase, they participated in a structured training phase that included guideline familiarisation, collective discussion of example cases, and pilot annotations on a small subset of the data. This pilot phase served to refine both the guidelines and the annotators' shared understanding of the categories to be applied.

Annotation was performed using *INCEpTION*[3], which supports span-level annotations with type (RF or PF) and feature assignments (exact labels of respective factors). All data was stored on an institute-owned server with restricted use and access only via the *INCEpTION* interface. Regular calibration meetings were held throughout the annotation process to discuss ambiguous cases, ensure consistency, and update the guidelines.

The dataset was divided into multiple subsets for annotation. Two annotators independently annotated one subset[4], and two different annotators independently annotated a second subset. A small portion of the data was annotated by all four annotators who took part in the second iteration to facilitate comprehensive reliability assessment. The task of the annotators was to (i) identify relevant spans in the conversations between therapist and client, (ii) decide whether the span was a protective or a risk factor and, finally, (iii) decide which factor exactly was represented, following the definitions in Table 2 and further hints from our annotation guidelines. Descriptions of CSAM use since the last session were annotated separately from the factor scheme as *Relapse Behaviour*. Additionally, two annotators labelled the entire session with a binary document-level label for relapse (*known relapse* or *no relapse* since the last session).

Across the 31 annotated sessions, annotators identified 1,670 factor instances in total, comprising 775 risk factors and 895 protective factors. This corresponds to an overall pooled protective ratio of 0.54, indicating a slight predominance of protective over risk-related client talk in the pilot corpus. We report the current status and results of annotations and corresponding challenges after two iterations of group discussion and guidelines as well as scheme refinement in the following. Further results and considerations are reported in Appendix A.

Inter-annotator agreement (IAA) was calculated on overlapping subsets of the data to monitor annotation reliability and to guide iterative revisions of the scheme. For this, we used *Krippendorff's* α *(unitizing)* (Krippendorff et al., 2016) computed natively in *INCEpTION*, giving credit for partial overlap. To analyse IAA in more detail, we also computed *relaxed* F_1[5] (Hripcsak and Rothschild, 2005). We performed label-aware 1:1 greedy maximum-overlap matching (>0 character overlap) in three modes (*general*, *exact*, *categorised*). True positives were counted as overlap and agreement on *exact* factor label, *general* type (RF/PF), or *category*. We aggregated results as pooled micro-F_1, and session-macro. Differences across iterations and between *exact* vs. *categorised* were quantified with bootstrap percentile 95% CI.

	i_1	i_2	$\Delta(i_2 - i_1)$
$\langle pos \rangle$	0.33 [0.22–0.41]	0.60 [0.53–0.67]	0.27*** [0.15–0.40]
RF	0.15 [0.02–0.26]	0.36 [0.20–0.54]	0.22† [0.02–0.42]
PF	0.33 [0.25–0.41]	0.40 [0.27–0.54]	0.07 (n.s.) [-0.09–0.23]

Table 3: α_u. Improvement over iterations (i) per annotation dimension: $\langle pos \rangle$ for position, *RF* and *PF* for risk factor and protective factor, respectively. Means; 95% CI in brackets. Significance of Δ tested via permutation tests: $^{***}p < .001$, $^{\dagger}p < .10$, n.s. $=$ not significant.

Krippendorff's α *(unitizing)* agreement on the span-level *position* of annotations increased significantly across iterations (details in Table 3). Agreement on *risk factors* improved descriptively, approaching conventional significance. For *protective factors*, no robust improvement was observed. On one session annotated by four annotators within the second iteration, α_u scores reached 0.48 for position, 0.47 on risk factors and 0.40 on protective factors, respectively.

Session-macro micro- and macro-F_1 improved significantly across sessions at both the *general* (RF vs. PF) and the *exact* factor level (see Table 4). These gains indicate that calibration meetings held between iterations improved both the coarse decision between risk or protective factor and the finer-grained code assignments, with the largest absolute improvements observed at the *general* level.

[3] https://inception-project.github.io/; version 35.2.

[4] One subset was annotated by three annotators. Normalisation was performed when computing IAA.

[5] Used libraries are reported in Appendix A.

	i_1	i_2	$\Delta(i_2 - i_1)$
micro $_{general}$	0.56 [0.47–0.64]	0.68 [0.60–0.75]	0.13*** [0.05–0.20]
micro $_{exact}$	0.41 [0.32–0.49]	0.53 [0.45–0.61]	0.12** [0.03–0.21]
macro $_{general}$	0.52 [0.43–0.60]	0.67 [0.58–0.75]	0.15*** [0.05–0.23]
macro $_{exact}$	0.31 [0.21–0.40]	0.43 [0.33–0.53]	0.12* [0.01–0.23]

Table 4: Relaxed F_1. Improvement over iterations (i): on the type level (*general*, RF vs. PF) and the *exact* factor level. The rows labelled *micro* and *macro* report session-macro micro-F_1 and session-macro macro-F_1, respectively. Point estimates; 95% CI in brackets. Δ: session-level bootstraps; significance tested with unpaired bootstrap: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.

Annotator confusion between factors and contrastive disagreements. In a subsequent effort to improve stability and annotator consistency, we tested grouping semantically similar fine-grained labels into higher-level categories within their type (RF and PF, respectively). Factors where no adjacent label was sufficiently similar to justify merging were retained as singleton categories. Pooling sessions from both iterations, agreement improved only in a modest, albeit consistent, manner.

In the second iteration of annotations, the main confusion hubs were risk factors of the category *Affect-/stress-driven coping and impulsivity*. Factors within *Self-regulation and functional coping* mirrored this on the protective side. Importantly, the factor with the most observed contrastive disagreements between risk and protective factors was *Recognition and functional handling of impulses* (PF), with relatively high counts of *confusions* with both *Dysfunctional coping (strategies)* (RF) and *Lack of impulse control* (RF).

6. Results

We present quantitative analyses of annotated risk and protective factors, tracking their distribution across sessions and around relapse, and report pilot span-extraction results with large language models (LLMs). Further results are reported in Appendix B.

6.1. *Protective Ratio* Over Time

To examine changes in the relative distribution of protective versus risk factors across sessions, we computed a *Protective Ratio (PR)* per client and session, $n_{PF}/(n_{PF} + n_{RF})$, where n_{PF} and n_{RF} are the numbers of annotated protective and risk factors.

At the session level, the *Protective Ratio* had a mean of 0.55, a median of 0.52, and ranged from 0.12 to 1.00.

A linear mixed-effects model revealed a significant positive effect of session number ($\beta = 0.11$, $SE = 0.03$, $p < .001$, 95% CI [0.05, 0.16]). This indicates that the relative proportion of protective factors systematically increased over time across all clients.

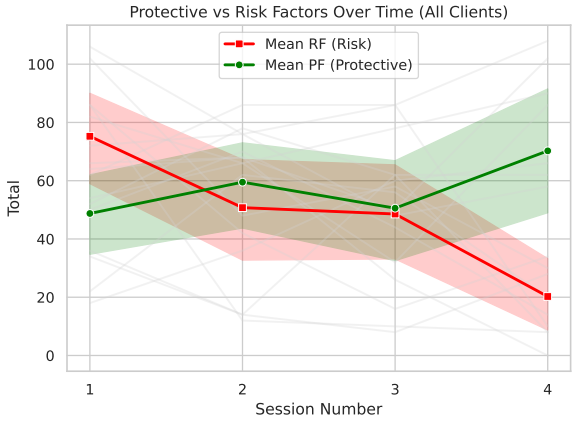


Figure 2: The trends of risk (red) and protective (green) factors over time, i.e., intervention sessions.

We also plotted mean per-client counts of protective and risk factors by session number to visualise their trajectories (Figure 2). In addition, we calculated monotonic trend tests for each individual client. While some clients showed clear monotonic increases in *PR*, others displayed more variable trajectories.

6.2. *Protective Ratio* Around Relapse

To examine whether the balance between protective and risk factors changes around relapse, we aligned sessions to self-reported relapses (pre- or post-relapse session windows) and contrasted them against all other sessions from the same client (baseline) using two-sided permutation tests. Differences (Δ) are computed as window minus baseline; negative values indicate a decrease from baseline. For the *Protective Ratio*, we summarise both a pooled, and the mean session-wise *PR*. In the pre-relapse window, both pooled and mean *PR* decreased relative to baseline (Δ = -0.10). In the post-relapse window, pooled *PR* was likewise lower (Δ = -0.07) and the mean *PR* showed the largest drop (Δ = -0.16). Though these effects are exploratory at the present pilot size, sessions immediately surrounding relapse contain a relatively higher share of risk than protective factors.

6.3. Factor-Level Analyses Around Relapse

For analyses on the *exact* factor level, we summarise the largest directional movers as exploratory effect sizes. We quantified relapse-aligned changes for each exact factor using two lenses: *presence* (difference in the proportion of sessions in which a factor appears), and normalised *rate* (difference in factor counts divided by total annotations in the respective session), indicating the relative dominance of these factors just before or after relapse.

For pre-relapse sessions, across lenses, the strongest changes were increases for mentions of *Avoidance, suppression, missing acceptance of the sexual preference* (RF3; $\Delta_{presence}$ = 0.52, Δ_{rate} = 0.05) and *Comorbidities, low well-being* (RF7; $\Delta_{presence}$ = 0.41, Δ_{rate} = 0.05), and smaller decreases across several protective factors. Post-relapse, *Cooperation and commitment* (PF8), *Skills to satisfy sexual needs in a healthy way* (PF10), *Recognition of the abuse of children* (PF5), and *Healthy sexual behaviour* (PF6) appeared less often, whereas *Dysfunctional coping* (RF4) and *Avoidance and non-acceptance of the sexual preference* (RF3) were comparatively more dominant. An across-client sensitivity analysis (global label shuffling) yielded similar rankings.

6.4. LLM-based Span Extraction

To explore the feasibility of automatic span extraction for risk and protective factors in CSAM-related therapeutic chats, we conducted pilot experiments using expert-annotated sessions as gold standard data.

To evaluate large language models (LLMs), we used *Qwen2.5:14b*[6] and *Mistral:7b*[7] in few-shot and fine-tuned setups. Prompts were enriched with examples and definitions derived from our annotation guidelines. *Qwen2.5:14b* served as the primary model, with *Mistral:7b* as a secondary baseline. Classification was performed both at the span level (*span*), and for entire messages (*message*), with separate runs for risk and protective factors (best results reported in Table 5).

In addition, we experimented with three BERT-based models. While these models could capture some patterns, their performance was substantially lower than that of the LLMs. Overall, protective factors were more accurately detected than risk factors. Further details, prompts, and results of the experiments will be reported in Appendix C.

[6] last accessed on 04.10.2025 via https://ollama.com/library/qwen2.5

[7] last accessed on 04.10.2025 via https://ollama.com/library/mistral

Model	Type / Level	$\mathbf{F_1}$ Score
Qwen2.5:14b (ft)	PFs (span)	0.350
Qwen2.5:14b (fs)	RFs (message)	0.296

Table 5: Best results of LLM span extraction experiments. "ft" refers to fine-tuned models, while "fs" refers to few-shot models.

7. Discussion and Outlook

Improved agreement across two annotation iterations indicates that annotator calibration meetings and revised, specific guidelines translate into better convergence. While absolute agreement scores remain moderate, we consider this a good result for pilot annotation, considering the task's linguistic subtlety and the breadth of the taxonomy. Importantly, many observed *disagreements* appear to reflect genuine clinical ambivalence rather than noise*. Several factors are best understood on a continuum, moving toward risk or protection depending on how clients express themselves on the same topic*. A concrete example are relationships: *trustful, fulfilled intimacy* functions protectively, whereas *lack or poor quality of intimacy* could increase risk. Simply "being in a relationship" is not protective without evidence of quality*. Whether a segment is coded as risk or protective factor frequently hinges on span boundaries as well, and whether annotators actually privilege intention or recognising the importance of certain concepts over actual behaviour as per our guidelines. Consequently, most *contrastive disagreements* happen around *coping and impulse control*; conscious reflection of dysfunctional coping could itself be protective, yet is easy to misread as risk*. These factors were also the most disputed in calibration meetings*.

Complementary analyses suggest that the scheme captures clinically coherent dynamics*. The *Protective Ratio* shows a systematic increase across sessions of all clients, while relapse-aligned windows show lower PR (a higher share of risk factors) in sessions before and after relapse. Clinically, post-relapse sessions often focus on relapse processing, which raises the salience of risk-coded talk even as it serves a therapeutic goal*. The pre-relapse elevation of clients expressing *Avoidance or non-acceptance* (RF3) and mentioning *Comorbidities or low well-being* (RF7) is intriguing against heterogeneous client profiles*, and indicates a possibility to detect factors that generalise across clients as early warning signals, given a larger database. While exploratory at current power, these patterns are directionally consistent with clinical experience*.

In pilot experiments, LLMs outperformed BERT-based models on span extraction, but absolute scores remain modest. This is plausible given the

small corpus, fine-grained labels, and the prevalence of mixed-evidence spans, and it mirrors the ambiguity seen in human annotation. Protective factors were detected more accurately than risk factors, which is consistent with their lower conceptual complexity and the smaller number of class distinctions. This pattern is also visible in IAA results, especially in the first annotation iteration.

Next, we will prioritise corpus growth, guideline tightening, and annotator training. Concretely, we will use fictional examples to synthesise diverse client profiles and intervention sessions. We will run adjudication sprints, and update the guidelines accordingly. Resulting annotations will be used for further annotator calibration, with a focus on span boundaries and contrastive risk versus protective factor decisions. To lower annotator cognitive load and improve label consistency, we suggest developing a flow-based decision aid that supports the coding of ambiguous spans and delivers tie-break rules. Future work could also examine client profiles to test whether specific constellations of factors are associated with elevated relapse risk, and further analyse criterion validity of the factors against relapse behaviour and therapist-judged relapse risk.

8. Conclusion

This work translates clinically meaningful risk and protective factors into span-level labels for therapist-led, prevention-oriented chats with voluntarily help-seeking individuals concerned about their risk of CSAM use. We introduce an annotation scheme and guidelines, ran a two-iteration pilot with targeted calibration, and observed reliability gains. Factor dynamics were examined using within-client baselines aligned to self-reported relapse. The *Protective Ratio* rises generally across sessions, but dips around relapse. Thus, the scheme is sensitive to session progression and relapse proximity and, with further guideline refinements and annotator training, can be used to create an actionable dataset. Although more annotated training and evaluation data are required before robust models are feasible, LLMs hold promise for identifying clinically relevant language patterns in prevention-oriented CSAM interventions.

Acknowledgements

We gratefully acknowledge funding from the German Federal Ministry of Research, Technology and Space (BMFTR) through the project VERANDA (16KIS2046K) and through the grant BIFOLD26B.

Limitations

We acknowledge several limitations of the present study. The current corpus is relatively small, which constrains the statistical robustness of our observations and limits the generalisability of the findings. Accordingly, some of the reported tendencies should be read as preliminary rather than conclusive. While inter-annotator agreement (IAA) remains comparatively low, we view this as an informative result in its own right. It reflects the conceptual and linguistic difficulty of coding psychosocial risk and protective factors in therapeutic discourse, where genuine ambivalence and span-boundary judgements are common. Model performance remains modest, reflecting both the small data size and the challenging, context-dependent nature of the task.

The dataset itself cannot yet be shared, as full anonymisation and completion of the annotation process are still ongoing. This necessary restriction currently limits reproducibility and external validation. However, we open-source non-sensitive artefacts related to this study, such as annotation guidelines, the label scheme, and results of our analyses.

Our scheme includes factors adapted from tools developed for forensic contexts. They are clinically plausible in our prevention-oriented *Dunkelfeld* chats, but not directly transferable. Base rates, help-seeking and disclosure incentives, and our span-level, language-based coding differ from file- or clinician-rated risk assessment. We therefore read our results with caution, as useful clinical cues rather than actuarial indicators. Future work should further establish validity in prevention cohorts.

Finally, potential sampling biases must be considered: the available data involve exclusively male clients, which may not fully represent the entire population of individuals seeking help for CSAM-related concerns. Despite these limitations, the study establishes a valuable foundation for future research on the linguistic identification of psychosocial risk and protective factors in prevention-oriented interventions.

Ethical Considerations

Working with data related to child sexual abuse material and associated risk/protective factors involves significant ethical and societal responsibility. The individuals whose data inform this work are part of a highly vulnerable and stigmatised population. All data were collected and processed in strict compliance with ethical guidelines and applicable data protection regulations. Prior to any analysis, data were de-identified and will undergo full anonymisation before potential publication or sharing. No

personally identifying information is disclosed at any stage of the project.

Due to the sensitive nature of the material, particular attention was paid to ensuring psychological safety for all researchers and practitioners involved in data handling and annotation. Regular open discussion was offered to minimise exposure-related distress and to maintain a high standard of ethical awareness.

Despite these challenges, this work addresses an ethically crucial domain: prevention and early intervention. Developing tools that can help detect psychosocial risk and protective factors in anonymised therapeutic communication has the potential to support therapists in their decision-making, contribute to offender prevention, and ultimately enhance victim protection. The overarching goal of this research is not surveillance or control, but the empowerment and support of therapeutic professionals and the advancement of evidence-based prevention efforts.

Ethical Approval. The chat study was reviewed and approved by the institutional ethics board of Charité – Universitätsmedizin Berlin under approval number EA4/187/24, ensuring compliance with all relevant ethical and legal standards for research involving sensitive and high-risk populations.

9. Bibliographical References

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders*, dsm-5-tr edition. American Psychiatric Association Publishing.

Paul C. Amrhein, William R. Miller, Carolina E. Yahne, Michael Palmer, and Laura Fulcher. 2003. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of Consulting and Clinical Psychology*, 71(5):862–878.

Kelly M. Babchishin, Ségolène Dibayula, Chiara McCulloch, R. Karl Hanson, and L. Maaike Helmus. 2023. ACUTE-2007 and STABLE-2007 predict recidivism for men adjudicated for child sexual exploitation material offending. *Law and Human Behavior*, 47(5):606–618.

Kelly M. Babchishin, R. Karl Hanson, and Heather VanZuylen. 2015. Online Child Pornography Offenders are Different: A Meta-analysis of the Characteristics of Online and Offline Sex Offenders Against Children. *Archives of Sexual Behavior*, 44(1):45–66.

Ibrahim Baroud, Lisa Raithel, Sebastian Möller, and Roland Roller. 2025. Beyond De-Identification: A Structured Approach for Defining and Detecting Indirect Identifiers in Medical Texts. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 75–85, Albuquerque, New Mexico. Association for Computational Linguistics.

Ricardo Barroso, Sofia Silva, Mariam Fishere, Julia Nentzl, Thuy Nguyen Vo, Carlos García Forero, Esperanza Luísa Gómez-Durán, Catarina Braz Ferreira, Hannes Gieseler, Viola Westfal, Berta Franch-Martínez, Lucie Krejčová, and Klaus M Beier. 2026. Child sexual abuse material (CSAM): a systematic review of risk profiles, risk factors, and typologies of users. *Sexual Medicine Reviews*, 14(1):qeaf081.

Ross M. Bartels and Hannah L. Merdian. 2016. The implicit theories of child sexual exploitation material users: An initial conceptualization. *Aggression and Violent Behavior*, 26:16–25.

Serra Baskurt, Kelly M. Babchishin, Gabriella Hilkes, and Michael C. Seto. 2025. A meta-analysis of recidivism rates among individuals who commit child sexual exploitation material (CSEM) offending. *Aggression and Violent Behavior*, 85:102080.

Klaus M. Beier, Dorit Grundmann, Laura F. Kuhle, Gerold Scherner, Anna Konrad, and Till Amelung. 2015. The German Dunkelfeld Project: A Pilot Study to Prevent Child Sexual Abuse and the Use of Child Abusive Images. *The Journal of Sexual Medicine*, 12(2):529–542.

Neha Deshpande, Mariam Fishere, and Stefan Hillmann. 2025a. The Development of an AI-Assistant to Therapists in a Chat-based Psychological Intervention: Gathering Users' First Impressions of the Experience. Cagliari, Italy.

Neha Deshpande, Stefan Hillmann, and Sebastian Möller. 2025b. Evaluating Large Language Models for Enhancing Live Chat Therapy: A Comparative Study with Psychotherapists. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 800–812, Avignon, France. Association for Computational Linguistics.

Angela W. Eke, L. Maaike Helmus, and Michael C. Seto. 2019. A Validation Study of the Child Pornography Offender Risk Tool (CPORT). *Sexual Abuse*, 31(4):456–476.

M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz, and A. D. Blackwell. 2021. Understanding the relationship between patient lan-

guage and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3):300–312.

Stephen D. Hart and Douglas P. Boer. 2020. Structured professional judgment guidelines for sexual violence risk assessment: the sexual violence risk-20 (SVR-20) versions 1 and 2 and risk for sexual violence protocol (RSVP). In *Handbook of violence risk assessment*, pages 322–358. Routledge.

L. Maaike Helmus, Angela W. Eke, and Michael C. Seto. 2025. What risk assessment tools can be used with men convicted of child sexual exploitation material offenses? Recommendations from a review of current research. *Law and Human Behavior*, 49(1):71–88.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association: JAMIA*, 12(3):296–298.

Tegan Insoll, Valeriia Soloveva, Eva Díaz Bethencourt, Anna Katariina Ovaska, Juha Nurmi, Arttu Paju, Mikko Aaltonen, and Nina Vaaranen-Valkonen. 2024. Factors Associated with Help-Seeking Among Online Child Sexual Abuse Material Offenders: Results of an Anonymous Survey on the Dark Web. *Journal of Online Trust and Safety*, 2(4).

Sara Jahnke. 2018. The Stigma of Pedophilia: Clinical and Forensic Implications. *European Psychologist*, 23(2):144–153.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential Privacy in Natural Language Processing The Story So Far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

Christina Lohr, Franz Matthies, Jakob Faller, Luise Modersohn, Andrea Riedel, Udo Hahn, Rebekka Kiser, Martin Boeker, and Frank Meineke. 2024. De-Identifying GRASCCO – A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus. In Rainer Röhrig, Niels Grabe, Ursula Hertha Hübner, Klaus Jung, Ulrich Sax, Carsten Oliver Schmidt, Martin Sedlmayr, and Antonia Zapf, editors, *Studies in Health Technology and Informatics*. IOS Press.

Johanna Lätth, Valdemar Landgren, Allison McMahan, Charlotte Sparre, Julia Eriksson, Kinda Malki, Elin Söderquist, Katarina Görts Öberg, Alexander Rozental, Gerhard Andersson, Viktor Kaldo, Niklas Långström, and Christoffer Rahm. 2022. Effects of internet-delivered cognitive behavioral therapy on use of child sexual abuse material: A randomized placebo-controlled trial on the Darknet. *Internet Interventions*, 30:100590.

Molly Magill, Timothy R. Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca E. F. Gordon, J. Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology*, 86(2):140–157.

Neil M. Malamuth, Christopher L. Heavey, and Daniel Linz. 1996. The Confluence Model of Sexual Aggression: Combining Hostile Masculinity and Impersonal Sex. *Journal of Offender Rehabilitation*, 23(3-4):13–37.

Aishik Mandal, Prottay Kumar Adhikary, Hiba Arnaout, Iryna Gurevych, and Tanmoy Chakraborty. 2025. A Comprehensive Survey of Datasets for Clinical Mental Health AI Systems. Version Number: 2.

Ruth E. Mann, R. Karl Hanson, and David Thornton. 2010. Assessing Risk for Sexual Recidivism: Some Proposals on the Nature of Psychologically Meaningful Risk Factors. *Sexual Abuse*, 22(2):191–217.

G. Alan Marlatt and Dennis M. Donovan. 2005. *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors*. Guilford press.

Vincent Marziano, Tony Ward, Anthony R. Beech, and Philippa Pattison. 2006. Identification of five fundamental implicit theories underlying cognitive distortions in child abusers: A preliminary study. *Psychology, Crime & Law*, 12(1):97–105.

William R. Miller. 2023. The evolution of motivational interviewing. *Behavioural and Cognitive Psychotherapy*, 51(6):616–632.

Amy Phenix, Yolanda Fernandez, Andrew JR Harris, Maaike Helmus, R. Karl Hanson, and David Thornton. 2017. *Static-99R coding rules, revised-2016*. Public Safety Canada.

Miriam Schuler, Hannes Gieseler, Katharina W. Schweder, Maximilian Von Heyden, and Klaus M. Beier. 2021. Characteristics of the Users of Troubled Desire, a Web-Based Self-management App for Individuals With Sexual Interest in Children: Descriptive Analysis of Self-assessment Data. *JMIR Mental Health*, 8(2):e22277.

Michael C. Seto and Angela W. Eke. 2015. Predicting recidivism among adult male child pornography offenders: Development of the Child Pornography Offender Risk Tool (CPORT). *Law and Human Behavior*, 39(4):416–429.

David Thornton, Gwenda M. Willis, and Sharon Kelley. 2024. Dynamic Protective Factors Relevant to Sexual Offending. *Current Psychiatry Reports*, 26(4):142–150.

Fritjof Von Franqué, Ralf Bergner-Koether, Stefanie Schmidt, Jan S. Pellowski, Jan H. Peters, Göran Hajak, and Peer Briken. 2023. Individuals under voluntary treatment with sexual interest in minors: what risk do they pose? *Frontiers in Psychiatry*, 14:1277225.

Tony Ward, Gwenda M. Willis, David S. Prescott, Stijn Vandevelde, Mary Barnao, and Wouter Wanzeele. 2025. *The Good Lives Model of Correctional Rehabilitation: Integrating Theory, Research, and Practice*. Advances in Preventing and Treating Violence and Aggression. Springer Nature Switzerland, Cham.

Gwenda M. Willis and Tony Ward. 2024. Evidence for the Good Lives Model in Supporting Rehabilitation and Desistance from Offending. In Leam A. Craig, Louise Dixon, and Theresa A. Gannon, editors, *The Wiley Handbook of What Works in Correctional Rehabilitation*, 1 edition, pages 299–309. Wiley.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet*, 15(3):110.

A. Annotation and Agreement

A.1. Consideration of a Scalar Factor Representation

While developing the annotation scheme and guidelines, we explored whether factors could be represented on a scalar continuum rather than as separate risk and protective categories. We decided against this in the present scheme because the relation between risk and protective factors is not fully symmetrical; some protective factors do not have a clear risk-side counterpart, and vice versa. A continuum-based formulation may nevertheless be worth revisiting in future work for selected factor domains.

A.2. Calculation of F_1

The F_1 computation pipeline was implemented in Python 3.10.8. We used dkpro-cassis 0.7.3 for reading INCEpTION XMI files and the corresponding type system, PyYAML 6.0.2 for configuration parsing, and numpy 1.26.4, pandas 2.2.2, and scipy 1.11.4 for metric computation, aggregation, bootstrap-based statistical analysis, and CSV export. F_1 was computed as a greedy span-overlap F_1 with 1:1 matching based on any character overlap of at least one character. We report three variants: *general* (risk vs. protective type only), *exact* (fine-grained factor labels), and *clustered* (fine-grained labels mapped to broader clusters). Scores were aggregated at multiple levels. We computed pairwise F_1 scores within each session and then summarised results as pooled micro-F_1, macro-F_1 across sessions, and macro-F_1 across labels, depending on the analysis. To compare annotation iterations, we estimated differences using session-level bootstrap confidence intervals with 5,000 resamples.

A.3. Exact Factors

Risk factor annotations were dominated by RF7, RF3, RF1, RF4, and RF4.1, while protective factor annotations were most frequent for PF4, PF2, PF8.1, PF4.1, and PF8. Table 6 shows the five most frequent labels per factor type. Table 7 and Table 8 report per-label relaxed *exact* F_1.

A.4. Clustering Factors into Categories

We considered clustering semantically similar fine-grained factors into categories, within their type. For risk factors, we distinguished between *Minor-focused sexual interest and hypersexuality*; *Offence-supportive cognitions and minimisation*; *Affect-/stress-driven coping and impulsivity*; *Intimacy or relationship deficits and social isolation*; *Prior behaviour and learning history*; *Structural instability and young age*; and *Avoidance and/or shame regarding the preference*. Analogously, for protective factors we defined the categories *Sexual responsiveness to adults and non-exclusivity*; *Accurate cognitive appraisal of CSA/CSAM and victim empathy*; *Self-regulation and functional coping*; *Intimacy, social embeddedness and stable relationships*; *Healthy sexual behaviour*; *Structural stability*, understood as satisfaction with housing and employment; and *Acceptance of the sexual preference*. Table 9 shows the clustering of semantically similar fine-grained labels into higher-level categories within their type (RF and PF, respectively). Table 10 reports overall relaxed F_1 for *exact* factors, *general* factor type, and *clustered* categories, indicating

	RF7	RF3	RF1	RF4	RF4.1	PF4	PF2	PF8.1	PF4.1	PF8
$n =$	159	97	85	82	66	163	123	121	119	119

Table 6: Top five most frequent risk and protective factor labels across the pilot corpus.

	PF1	PF2	PF3	PF4	PF4.1	PF5	PF6	PF6.1	PF7	PF8	PF8.1	PF10
i_1	0.72	0.49	0.37	0.58	0.22	0.49	0.07	0.67	0.44	0.25	0.36	0.42
i_2	0.60	0.72	0.29	0.61	0.47	0.44	0.50	0.67	0.80	0.51	0.46	0.00
$\Delta(i_2 - i_1)$	-0.12	0.23	-0.08	0.03	0.25	-0.04	0.43	0.00	0.36	0.26	0.10	-0.42
Support (i_1)	23.5	65.5	24.5	81.0	31.5	22.5	14.5	4.5	20.5	63.0	49.5	12.0
Support (i_2)	15.0	23.5	10.5	33.0	38.0	4.5	2.0	1.5	12.5	19.5	28.0	1.0

Table 7: Per-label relaxed *exact* F_1 for protective factors. Support is the average of label counts across the compared annotations within each iteration.

that categorising factors only slightly improved inter-annotator agreement.

A.5. Confusion and Contrastive Disagreements

We inspected pairwise confusions between fine-grained factor labels. These were concentrated in a small number of conceptually adjacent factors rather than spread broadly across the inventory. The most frequent confusion was PF4.1 vs. RF4 ($n = 5$), followed by PF4 vs. PF4.1, PF4.1 vs. PF8, PF4.1 vs. RF4.1, and RF4 vs. RF7 (each $n = 4$). Overall, the pattern suggests that residual disagreement was primarily contrastive, especially in passages concerning coping and impulse handling. Table 11 shows the most frequent confusions with $n \geq 2$. To further analyse where agreement breaks down, future work could also consider intra-annotator agreement, i.e., the extent to which individual annotators apply the scheme consistently across repeated passes. This would provide additional insight into whether disagreement stems primarily from annotator-specific variation or from ambiguities in the scheme itself.

B. Relapse-Aligned Analyses

B.1. *Protective Ratio* Around Relapse

To supplement the relapse-aligned results reported in Section 6.2, Table 12 provides the corresponding Protective Ratio (PR) values for the pre- and post-relapse windows and their within-client baselines. Sessions were aligned to self-reported relapse, using the session immediately before relapse as the pre-relapse window ($t = -1$) and the session immediately after relapse as the post-relapse window ($t = +1$). Baseline consisted of all other sessions from the same clients with listed relapse events. Differences (Δ) are computed as window minus baseline. For PR, we report both pooled PR and mean session-wise PR. Permutation tests were computed within client.

B.2. Factor-Level Analyses Around Relapse

Tables 13 and 14 list the top five exact-factor movers by absolute effect size for the pre- and post-relapse windows, respectively. For each factor, we report results under two complementary lenses: *normalised rate*, defined as the factor count divided by the total number of annotations in the respective session set, and *presence*, defined as the proportion of sessions in which the factor occurs at least once. All contrasts are computed as window minus baseline. Results are based on within-client permutation tests. The two panels in each table are ranked independently by absolute effect size and are therefore not row-wise matched.

C. Automated Span Extraction

C.1. LLM-based Results

For LLM-based extraction, we experimented with span-based and message-based detection and extraction of protective and risk factors using different LLMs, but achieving the best results with $Qwen2.5{:}14b$[8]. Underlined scores are the ones reported in the main paper.

Table 15 and Table 16 show the detailed results of detecting and classifying protective and risk factors, respectively, without fine-tuning the models. Fine-tuning the model on silver-annotated data for classification improved extracting protective factors, but worsened the extraction of risk factors. Table 17 and Table 18 show the results using the fine-tuned classifier model.

Finally, we also experimented with message-based classification, achieving the best result for

[8] last accessed on 04.10.2025 via https://ollama.com/library/qwen2.5

	RF0	RF1	RF1.1	RF2	RF2.1	RF3	RF4	RF4.1	RF5	RF6	RF7	RF8	RF8.1	RF9	RF10	RF10.1	RF11	RF12
i_1	0.00	0.52	0.00	0.48	0.29	0.47	0.26	0.43	0.57	0.47	0.36	0.29	0.22	0.00	0.62	0.20	0.00	0.50
i_2	0.00	0.69	0.89	0.67	–	0.72	0.36	0.44	0.00	0.40	0.66	0.00	0.40	0.33	0.33	0.17	–	0.00
$\Delta(i_2 - i_1)$	0.00	0.17	0.89	0.19	–	0.25	0.10	0.01	-0.57	-0.07	0.30	-0.29	0.18	0.33	-0.28	-0.04	–	-0.50
Support (i_1)	3.5	53.5	0.5	12.5	24.0	38.5	30.5	34.5	3.5	27.5	56.0	3.5	4.5	2.5	13.0	24.5	1.0	8.0
Support (i_2)	4.5	13.0	4.5	12.0	–	25.0	19.5	9.0	1.5	15.0	48.5	1.5	5.0	6.0	6.0	6.0	–	1.5

Table 8: Per-label relaxed exact F_1 for risk factors. Cells marked – indicate that the label was not present in that iteration. Support is the average of label counts across the compared annotations within each iteration.

	Category	Factors
PF_A	Sexual responsiveness to adults and non-exclusivity	PF1
PF_B	Accurate cognitive appraisal of CSA/CSAM and victim empathy	PF5
PF_C	Self-regulation and functional coping	PF4, PF4.1
PF_D	Intimacy, social embeddedness and stable relationships	PF2, PF7
PF_E	Healthy sexual behaviour	PF10, PF6, PF6.1
PF_F	Structural stability	
PF_G	Acceptance of the sexual preference	PF3
RF_A	Minor-focused sexual interest and hypersexuality	RF1, RF1.1, RF10, RF12
RF_B	Offence-supportive cognitions and minimisation	RF5, RF8.1, RF11
RF_C	Affect-/stress-driven coping and impulsivity	RF10.1, RF4, RF4.1, RF7
RF_D	Intimacy or relationship deficits and social isolation	RF2, RF2.1
RF_E	Prior behaviour and learning history	RF6
RF_F	Structural instability and young age	RF9
RF_G	Avoidance and/or shame regarding the preference	RF3

Table 9: Clustering of fine-grained factors to categories. Some factors remain as singleton categories.

risk factor extraction with this method. Detailed results are shown in Table 19 and Table 20.

The figures in Appendix C.1 show the respective prompts for span and message detection and classification.

C.2. BERT-based Span Extraction (Token Classification)

A list of the BERT models used for experiments can be found in Table 22. The results are presented in Table 21.

	pooled	session-macro micro	session-macro macro
exact	0.45	0.47	0.37
clustered	0.49	0.51	0.40
general	0.59	0.62	0.60

Table 10: Overall relaxed F_1 across all 31 sessions.

Factor A	Factor B	n
PF4.1	RF4	5
PF4	PF4.1	4
PF4.1	PF8	4
PF4.1	RF4.1	4
RF4	RF7	4
RF2	RF7	3
RF0	RF6	3
RF4	RF4.1	2
RF10.1	RF6	2
RF4	RF8.1	2
RF2	RF8.1	2
RF3	RF7	2
PF4	PF8	2
PF2	PF4	2
PF4.1	RF7	2

Table 11: Most frequent pairwise factor confusions. Only confusions with $n \geq 2$ are shown.

Window	Aggregation	PR_{win}	PR_{base}	Δ	p_{perm}
pre ($t = -1$)	pooled	0.439	0.539	-0.101	0.443
pre ($t = -1$)	mean	0.439	0.541	-0.102	0.545
post ($t = +1$)	pooled	0.454	0.521	-0.067	0.640
post ($t = +1$)	mean	0.407	0.565	-0.158	0.317

Table 12: Protective Ratio (PR) in pre- and post-relapse windows compared to within-client baseline. Δ is computed as window minus baseline.

normalised rate			presence		
Factor	Δ	p_{perm}	Factor	Δ	p_{perm}
RF7	0.053	0.157	RF3	0.524	0.266
RF3	0.049	0.321	RF10	-0.413	0.119
PF2	-0.044	0.041	RF7	0.413	0.270
RF1	0.037	0.246	RF1.1	-0.333	0.376
PF4	0.032	0.377	RF8.1	0.317	0.443

Table 13: Top five exact-factor movers by absolute effect size in the **pre-relapse window** ($t = -1$) under the within-client permutation scheme, shown separately for the normalised rate and presence lenses. The two panels are ranked independently by absolute effect size and are therefore not row-wise matched.

normalised rate			presence		
Factor	Δ	p_{perm}	Factor	Δ	p_{perm}
RF4	0.059	0.044	PF8	-0.571	0.058
RF3	0.054	0.286	PF10	-0.413	0.380
PF4	0.047	0.176	PF5	-0.381	0.288
PF4.1	-0.042	0.431	PF6	-0.381	0.466
RF6	-0.041	0.165	RF4	0.333	0.348

Table 14: Top five exact-factor movers by absolute effect size in the **post-relapse window** ($t = +1$) under the within-client permutation scheme, shown separately for the normalised rate and presence lenses. The two panels are ranked independently by absolute effect size and are therefore not row-wise matched.

metric	relaxed match	exact match
precision	0.484	0.481
recall	0.238	0.210
F1	**0.319**	**0.292**

Table 15: Evaluation for span detection and classification on protective factors with $Qwen2.5{:}14b$.

metric	relaxed match	exact match
precision	0.327	0.327
recall	0.149	0.149
F1	**0.205**	**0.205**

Table 16: Evaluation for span detection and classification on risk factors with $Qwen2.5{:}14b$.

metric	relaxed match	exact match
precision	0.431	0.439
recall	0.302	0.290
F1	**0.349**	**0.350**

Table 17: Evaluation for span detection and classification on protective factors with fine-tuned $Qwen2.5{:}14b$.

metric	relaxed match	exact match
precision	0.133	0.132
recall	0.158	0.149
F1	**0.145**	**0.140**

Table 18: Evaluation for span detection and classification on risk factors with fine-tuned $Qwen2.5{:}14b$.

metric	score
precision	0.230
recall	0.383
F1	**0.287**

Table 19: Evaluation for message classification on protective factors with $Qwen2.5{:}14b$.

metric	score
precision	0.272
recall	0.324
F1	**0.296**

Table 20: Evaluation for message classification on risk factors with $Qwen2.5{:}14b$.

Prompt for span detection of protective factors

```
Given the following message, extract text spans that could be potential protec-
tive factors in a therapeutical context.

Instruction:
  • Given the input message, output only the relevant text spans as string,
    nothing else.
  • If multiple Spans can be found, create a list of strings like this:
    ["span1", "span2"]
  • If no span is relevant, output exactly: None
  • Do not add explanations, notes, greetings, or any extra words

Example 1:
input: I can imagine the struggle. Personally, I am finding it rather easy
to engage in a conversation with you
output: I am finding it rather easy to engage in a conversation with you
Example 2:
input: I also joined a choir, go to a theatre improvisation class, see a
friend for coffee every morning (which my wife has a hard time to accept) but
it helps me to gain more confidence in accepting myself
output: ["joined a choir, go to a theatre improvisation class, see a friend
for coffee every morning", "it helps me to gain more confidence in accepting
myself"]
Example 3:
input: Hi!
output: None
Now, classify the following message:
input: {message}
```

Figure 3: Prompt for span detection of protective factors.

Model	Metric	Protective Factors	Risk Factors
		Score	Score
bert-base-cased	Precision	0.002	0.0009
	Recall	0.033	0.010
	Accuracy	0.753	0.693
	F1	**0.004**	**0.0016**
DisorBERT	Precision	0.003	0.0005
	Recall	0.049	0.010
	Accuracy	0.719	0.524
	F1	**0.005**	**0.0009**
MentalBERT	Precision	0.004	0.001
	Recall	0.066	0.010
	Accuracy	0.713	0.738
	F1	**0.007**	**0.0022**

Table 21: Evaluation for token classification on protective and risk factors with fine-tuned BERT models.

Model name	Source	Last Accessed
bert-base-cased	https://huggingface.co/google-bert/bert-base-cased	04.10.2025
DisorBERT	https://huggingface.co/citiusLTL/DisorBERT	04.10.2025
MentalBERT	https://huggingface.co/mental/mental-bert-base-uncased	04.10.2025

Table 22: List of BERT models used for experiments.

Prompt for span detection of risk factors

Given the following message, extract text spans that could be potential risk factors in a therapeutical context.
Instruction:

- Given the input message, output only the relevant text spans as string, nothing else.
- If multiple Spans can be found, create a list of strings like this: ["span1", "span2"]
- If no span is relevant, output exactly: None
- Do not add explanations, notes, greetings, or any extra words

Example 1:
input: Good.. I saw on your questionnaire that you are not in a relationship, and that you are not happy about that. Did I get that right?
output: you are not in a relationship, and that you are not happy about that.

Example 2:
input: Im not good at speaking with other people, due to my anxiety problems 😟
output: ["Im not good at speaking with other people", "my anxiety problems"]

Example 3:
input: Hi!
output: None

Now, classify the following message:
input: {message}

Figure 4: Prompt for span detection of risk factors.

Prompt for span classification of protective factors

You are an annotation system. You must respond **only** with the relevant category labels, exactly as specified. Categories: {protective_factors}

Instruction:

- Given the input span, output only the relevant category labels (e.g., "PF2","PF7"), nothing else.
- If no category applies, output exactly: None
- Do not add explanations, notes, greetings, or any extra words
- One span may fit under multiple categories.

Example 1:
input: Personally, I am finding it rather easy to engage in a conversation with you
output: PF2

Example 2:
input: Just masturbation, sometimes i use legal porn or fantasies of minors (which im trying to cut down on) and i've started using <NAME> chatbots for roleplay - only involving adults
output: PF6,PF10

Example 3:
input: Hi!
output: None

Now, classify the following span:
input: {span}

Figure 5: Prompt for span classification of protective factors.

Prompt for span classification of risk factors

```
You are an annotation system.  You must respond **only** with the relevant cat-
egory labels, exactly as specified.
Categories:
{risk_factors}
Instruction:
  • Given the input span, output only the relevant category labels (e.g.,
    "RF2","RF7"), nothing else.
  • If no category applies, output exactly:  None
  • Do not add explanations, notes, greetings, or any extra words
  • One span may fit under multiple categories.
Example 1:
input:  you are not in a relationship, and that you are not happy about that.
output:  RF2
Example 2:
input:  Im not good at speaking with other people, due to my anxiety problems
output:  RF2,RF7
Example 3:
input:  Hi!
output:  None
Now, classify the following span:
input:  {span}
```

Figure 6: Prompt for span classification of risk factors.

Prompt for message classification of protective factors

```
You are an annotation system.  You must respond **only** with the relevant cat-
egory labels, exactly as specified.
Categories:
{protective_factors}
Instruction:
  • Given the input message, output only the relevant category labels (e.g.,
    "PF2","PF7"), nothing else.
  • If no category applies, output exactly:  None
  • Do not add explanations, notes, greetings, or any extra words
  • One message may fit under multiple categories.
Example 1:
input:  I can imagine the struggle.  Personally, I am finding it rather easy
to engage in a conversation with you
output:  PF2
Example 2:
input:  Just masturbation, sometimes i use legal porn or fantasies of minors
(which im trying to cut down on) and i've started using <NAME> chatbots for
roleplay - only involving adults
output:  PF6,PF10
Example 3:
input:  Hi!
output:  None
Now, classify the following message:
input:  {message}
```

Figure 7: Prompt for message classification of protective factors.

Prompt for message classification of risk factors

```
You are an annotation system.  You must respond **only** with the relevant cat-
egory labels, exactly as specified.
Categories:
{risk_factors}
Instruction:
  • Given the input message, output only the relevant category labels (e.g.,
    "RF2","RF7"), nothing else.
  • If no category applies, output exactly:  None
  • Do not add explanations, notes, greetings, or any extra words
  • One message may fit under multiple categories.
Example 1:
input:  Good..  I saw on your questionnaire that you are not in a relationship,
and that you are not happy about that.  Did I get that right?
output:  RF2
Example 2:
input:  Im not good at speaking with other people, due to my anxiety problems
😟
output:  RF2,RF7
Example 3:
input:  Hi!
output:  None
Now, classify the following message:
input:  {message}
```

Figure 8: Prompt for message classification of risk factors.

Automatic Detection of Direct and Self-Repetitions in Naturalistic Speech Recordings of French- and Dutch-Speaking Autistic Children

Federica Beccaria[1,2], **Marie Kolenberg**[2], **Pierre Labendzki**[7], **BeLAS Consortium**, **Inge Zink**[2,3], **Mikhail Kissine**[1,4,5]

[1]Autism in Context: Theory and Experiment (ACTE), Center for Linguistic Research (LaDisco), ULB Neuroscience Institute, Université Libre de Bruxelles

[2]Experimental Oto-Rhino-Laryngology (ExpORL), KU Leuven

[3]Leuven Autism Research (LAuRes), KU Leuven

[4]Department of Linguistics, University College London

[5]Department of Philosophy, Classics, History of Art and Ideas, University of Oslo

[7]University of East London

federica.beccaria@ulb.be, marie.kolenberg@kuleuven.be, labendzki@uel.ac.uk, inge.zink@kuleuven.be, m.kissine@ucl.ac.uk

Abstract

This study investigates the use of cosine similarity measures across syntactic, lexical, and semantic vector representations to detect repetitions in the spontaneous speech of autistic children. It focuses on direct repetitions (i.e., immediate verbatim repetitions of linguistic output produced by another individual) and self-repetitions (i.e., within-speaker recurrence). The performance of similarity-based methods is then compared with state-of-the-art black-box classification models based on BERT, trained on the same data. Using spontaneous speech data from French- and Dutch- speaking autistic children, the results show that lexical and semantic similarity provide reliable cues for identifying self-repetitions, achieving high precision and recall, with F1-scores exceeding 83%, comparable to those obtained by BERT-based models. In contrast, direct repetitions are more difficult to detect using similarity-based approaches, with BERT models clearly outperforming them and reaching F1-scores above 73%. Across all conditions, syntactic similarity consistently underperforms relative to lexical and semantic measures. These findings highlight the strengths and limitations of similarity-based approaches and suggest directions for future research, particularly in improving the detection of direct repetitions and assessing the cross-linguistic generalizability of these methods.

Keywords: Autism, Direct Repetitions, Self-Repetitions, Echolalia, Cosine Similarity, BERT, Repetition Detection

1. Introduction

Autism is a neurodevelopmental condition characterized by a wide range of developmental features, including differences in social communication and repetitive behavior patterns (American Psychiatric Association, 2013; Schaeffer et al., 2023).

Echolalia, the repetition of previously heard speech, is often regarded as a core feature of autism due to its prevalence in the language of autistic individuals, with variation depending on language proficiency (Maes et al., 2024). However, definitions of the phenomenon vary widely, and the distinction between echolalia and repetitions observed in neurotypical language development is not clearly delineated. Traditionally, categories of echolalia differ both in their formal resemblance to the source segment (*pure* vs. *mitigated* echolalia) and in their timing relative to the source (*direct* vs. *delayed* echolalia, where the latter may also include sources from outside the conversation, such as songs). However, the definitions of these categories and their inclusion under the phenomenon of echolalia differ between authors. Similarly, self-repetitions have been considered (McFayden et al., 2022), or explicitly excluded (van Santen et al., 2013), as instances of echolalia, or rather as a related non-generative phenomenon, broadly defined as the reuse of previously produced or perceived linguistic material (Luyster et al., 2022). Some definitions exclude all repetitions that display communicative intent (e.g., questions for clarification) or that do not mimic the prosody of the source (Amiriparian et al., 2018; Marom et al., 2018), while others accept formal and functional variation (Pascual et al., 2017; Xie et al., 2023). This lack of consensus complicates systematic analyses, particularly in large language corpora, as definitions often rely on detailed pragmatic and conversational analyses to determine whether a linguistic segment qualifies as echolalia (Ryan et al., 2024).

Proceedings of the 6th RaPID@MENTAl.ai Workshop @ LREC 2026, pages 146–156
12 May 2026.

In this context, some researchers have attempted to develop methods to automatically extract segments of echolalic speech. Some approaches rely on acoustic analysis to examine spectral similarities between sentences (Amiriparian et al., 2018), while others focus on transcription-based analyses to identify repetitions (Bigi et al., 2014; van Santen et al., 2013). From this perspective, Fusaroli et al. (2023) have made significant contributions by reframing the study of echolalia through the lens of alignment theory. Their methodology involves computing alignment rates across linguistic representations of different types (syntactic, lexical, and semantic) between autistic children and their caregivers to quantify the degree of recycling language material. This approach offers valuable insights into the interactive dynamics of language in autism. Building on this foundation, our study adapts and extends Fusaroli et al. (2023)'s approach with a novel aim: instead of computing a global alignment or repetition rate, we seek to *detect* recurring linguistic units by comparing pairs of segments, contrasting those classified as repetitive with those classified as non-repetitive. By establishing thresholds for syntactic, lexical, and semantic similarity on an extensively annotated gold standard dataset, we enable an efficient and scalable approach for detecting repetitive speech. This approach facilitates a detailed analysis of echolalia, providing insights into its linguistic features, length, and communicative functions. Furthermore, the success of each similarity computation in detecting repetitive pairs informs us of the linguistic information (syntactic, lexical, and semantic) that leads listeners to perceive sameness in a source-echolalic pair. In a next step, we compare the results of these linguistically informed methods with those of state-of-the-art pretrained BERT models (Devlin et al., 2019), fine-tuned on our classification task.

2. Methods

The data used for the development of the models presented in this study were drawn from the Belgian Language in Autism Study (BeLAS). The sample comprises naturalistic speech recordings from 14 French- and 15 Dutch-speaking children aged between 2 and 6 years (mean = 55.81 months, SD = 10.66 months; 19 males, 10 females). All children had a formal autism diagnosis and were administered the Autism Diagnostic Observation Schedule - Second Edition (ADOS-2; Lord et al., 2012). Children were additionally assessed using standardized instruments to evaluate expressive and receptive language development quotients (Bayley Scales of Infant and Toddler Development, Clinical Evaluation of Language Fundamentals - Preschool, Evalo, Peabody Picture Vocabulary Test; Bayley, 2006; Semel et al., 2020; Ortho Édition, 2009; Schlichting, 2005; Dunn and Dunn, 2019) as well as non-verbal cognitive skills (Snijders-Oomen Non-verbal Intelligence Test; Tellegen and Laros, 2017). As shown in Table 1, no statistically significant differences were observed between French- and Dutch-speaking children on any of the reported measures, indicating that the two language groups were well matched. Importantly, however, participants exhibited substantial variability in both linguistic and non-verbal cognitive skills, allowing the sample to represent a broad range of developmental profiles within the autism spectrum. Speech recordings were collected over approximately six hours in the children's homes using a small lapel recorder placed in the pocket of a project-designed T-shirt. Then, the hour with the highest amount of each child's speech was selected using a pre-trained diarization model (Lavechin et al., 2021). Finally, we orthographically transcribed at least 20 minutes of child speech, with duration adjusted according to individual language output.

2.1. Gold Standard Annotation

To establish a gold standard annotation for the repetition detection task, we manually coded direct and self-repetitions in a total of 760 minutes of audio recordings. Each participant contributed at least 20 minutes of audio, with some providing up to 60 minutes depending on the amount of language produced. Of the total, 360 minutes were annotated for 14 French-speaking children and 400 minutes for 15 Dutch-speaking children. Coding was performed using Praat (Boersma and Weenink, 2025).

Direct repetitions were defined as linguistic units occurring within a maximum of 10 seconds of the source clause, sharing at least one content word irrespective of morphological changes. In example 1, produced by a Dutch-speaking child in our corpus, *tickle* is shared between the utterance of another speaker and the autistic child, appearing in two different morphological forms: the first in the third person singular and the second in the first person singular of the present tense.

(1) **Other Speaker:** *Kietelt* dat?
'Does that tickle?'
Autistic Child: *Kietel*
'Tickle'

Self-repetitions were defined as exact reiterations of segments from the child's own language productions. Among the three examples reported below, all produced by the same autistic child in our French-speaking sample, examples 2a and 2b

Measure	French (n = 14)	Dutch (n = 15)	t-value (df)
Age (months)	57.67 (8.45; 42.41–71.06)	55.81 (10.66; 40.01–71.12)	-0.52 (26.34)
ADOS CSS	6.14 (2.38; 2–10)	5.00 (1.60; 2–7)	-1.51 (22.58)
Non-verbal IQ	90.23 (20.82; 55–117)	92.67 (14.79; 57–115)	0.35 (21.30)
Expressive Language	77.27 (32.98; 26.80–117.53)	89.14 (19.53; 57.08–116.41)	0.88 (8.03)
Receptive Language	80.07 (27.11; 32.55–132.01)	89.19 (19.71; 56.68–116.89)	0.95 (17.39)

Table 1: Descriptive statistics of participants by language group (French vs. Dutch), including mean, standard deviation, and range. Independent-samples t-tests indicate no significant group differences

are considered self-repetitions, but 2a and 2c are not, since not all the material of 2a is repeated.

(2) a. **Autistic Child:** *c'est une voiture de police*
'It's a police car'
b. **Autistic Child:** *c'est une voiture de police*
'It's a police car'
c. **Autistic Child:** *c'est* la *police*
'It's the police'

For more information about the coding protocol for the gold standard, see this OSF repository. To assess the reliability of manual coding, 10% of all transcribed audio files were double-coded. Coders demonstrated very high agreement: for direct repetitions, agreement was 95.5% (Cohen's Kappa = 0.84), whereas for self-repetitions, agreement reached 99.4% (Kappa = 0.82), reflecting almost perfect consistency between coders.

2.2. Model Development for Repetition Detection

Since the recordings were obtained without explicit instructions or control over background noise, we opted against an audio-based approach for repetition detection. Instead, we developed a model based on orthographic transcriptions of speech produced by autistic children and other speakers, building on the methodology of Fusaroli et al. (2023) with adaptations for multiple languages and interlocutors. This framework was applied to both direct and self-repetitions.

A linguistic unit is any child's language production, regardless of syntactic complexity, ranging from single words to complete sentences. Non-linguistic vocalizations, such as babbling, were excluded. This inclusive definition allows representation of the full range of linguistic output, including children with limited verbal skills.

Following Fusaroli et al. (2023), each unit was represented using syntactic, lexical, and semantic vectors. Syntactic vectors encode sequences of Part-Of-Speech (POS) tags (e.g., nouns, verbs), lexical vectors encode sequences of lemmas (e.g., *child* for *children*; *write* for *wrote*), and semantic vectors provide a holistic representation of the unit's meaning. Together, these vectors capture the main linguistic information available from transcribed speech.

Cosine similarity was computed for each unit: direct repetitions with other speakers' units within 10 seconds, self-repetitions with the child's earlier productions. These pairs, together with gold-standard annotations, were then used to fine-tune French- and Dutch-language BERT models for repetition detection.

2.3. Cosine Similarity Models: Vector representation, Similarity Measures, and Performance Evaluation

For syntactic vectors, we used spaCy models *fr core news sm* for French and (*nl core news sm* for Dutch; Honnibal and Montani, 2017) to determine POS tags, grouped into n-grams with *n=2*, as per Fusaroli and colleagues' (2023) findings. Due to the large number of short linguistic segments (< 4 words), we opted against using larger n-grams. If a linguistic unit contained fewer tokens than the selected *n=2*, the entire one-word segment was treated as a single n-gram.

Similarly, we used spaCy's module *Lemmatizer* to create a list of unique lemmas. Then, for each file, we constructed a combined list containing all the unique lemmas and POS n-gram sequences. All linguistic units were then represented as single vectors, where each value indicated the number of times (0, 1, 2, etc.) each lemma or POS n-gram from the list appeared in the linguistic segment. This ensured uniform vector structure across speakers, facilitating meaningful comparisons regardless of the linguistic segment's length. Function words were included, as their proportional presence across the considered segments affected similarity measures minimally.

For semantic vectors, we employed BERT SentenceTransformers models trained on French (*CamemBERT* large, Martin et al.2020) and Dutch

(*RobBERT*, Delobelle et al. 2020). These models generated fixed-length embedding of 1024 dimensions for French and 768 for Dutch, aligning with the one-dimensional format supported by the Python SentenceTransformers library (Reimers and Gurevych 2019, 2020). Each dimension corresponds to a numerical feature encoding some aspect of textual meaning; the full set of dimensions jointly represents the text.

After constructing vector representations, cosine similarity scores were calculated using the Sentence Transformers *cos sim* function to compare pairs of linguistic segments. The autistic child's linguistic productions were compared to (i) those of other speakers that occurred at most 10 seconds earlier and (ii) all those they had previously produced (self-repetition).

Next, we aimed to determine which *cosine similarity thresholds* yielded the best results in distinguishing non-repetitive from repetitive production pairs. A range of 100 thresholds between -1 and 1 (corresponding to the cosine similarity function values) with a step size of 0.02 was tested for each measure, and the resulting precision and recall values were evaluated. Our goal was to maximize recall (i.e., the proportion of repetitions correctly detected) while maintaining precision (i.e., the proportion of predicted repetitive cases that were actually repetitive) at an acceptable level (Table 2). Finally, we evaluated the performance (precision, recall, F1-score) of the selected thresholds for each measure.

2.4. BERT Models: Train-Test Split, and Performance Evaluation

In this study, we further compared the performance of our repetition-detection approach with state-of-the-art BERT models. To this end, we fine-tuned BERT models for sequence classification, using the same base models as the SentenceTransformer models employed for constructing semantic vectors (*CamemBERT* large, Martin et al.2020 for French and *RobBERT*, Delobelle et al. 2020 for Dutch). Fine-tuning and evaluation involved creating training and test sets, with a speaker-based split so that the models were evaluated on speech from children not encountered during training. For a direct comparison with BERT, the cosine similarity models were evaluated on BERT's test set here, rather than on the entire dataset as previously.

We isolated 4 French and 4 Dutch speakers (out of the total of 29) in the test set, based on a 4-means clustering of their characteristics (age, expressive and receptive language development quotients) and the number of direct and self-repetitions that they produced. This ensured that similar language profiles of speakers were found in the train and test sets, and that the repetitive phenomena of interest occurred in a similar frequency in both.

The French and Dutch BERT models were then trained on the gold-standard annotations of the remaining 21 speakers for both direct and self-repetitions. Training was conducted using 10-fold cross-validation, with adjustments for unbalanced class weights (the repetitive class being underrepresented), and parameter evaluation based on the F1-score, with 'repetitive' as the positive class. Both BERT models were evaluated on the test set using precision, recall, and F1-score.

2.5. Materials

All code for computing similarity metrics, determining best similarity thresholds, making the train-test split, training the BERT models, evaluating the models, and creating visualizations is available in this OSF repository. The repository also contains details on the characteristics of the train and test sets and the parameters used for training the BERT models.

Moreover, this GitHub contains the finished models as well as Python code that allows users to try the models on their own data.

Data visualization was conducted using the Python libraries Seaborn (Waskom 2021) and Matplotlib (Hunter 2007). Generative AI tools were used to debug Python code (OpenAI 2025).

3. Results

This section presents the results for both direct and self-repetitions, comparing cosine similarities of syntactic, lexical, and semantic vectors across the French and Dutch datasets.

An additional analysis was conducted to compare these models with BERT-based models.

3.1. Performance of the Cosine Similarity Models

Figure 1 illustrates the overall performance of models based on syntactic, lexical, and semantic cosine similarities in distinguishing non-repetitive pairs from direct or self-repetitions. Receiver Operator Curves (ROC) in full lines plot the true positive against the false positive rate for the thresholds detecting direct repetitions. By contrast, dashed lines do so for the thresholds identifying self-repetitions. Overall, the Area Under the Curve (AUC) scores are quite satisfactory for all linguistic measures (above 73%), in both languages and phenomena. However, the ROCs are higher for self-repetitions than for direct repetitions across

the three measures. Secondly, AUC-scores are markedly lower for thresholds on syntactic similarity (73.2% and 76.2% for French and Dutch direct repetitions; 92.8% and 94.5% for Dutch and French self-repetitions) than for those on lexical and semantic similarity. Indeed, the latter scores between 88.6% for direct repetition and 99.9% for self-repetition. Lastly, performances of the thresholds on Dutch data are generally slightly lower than those of models on French data. In sum, the best-performing models are those that detect self-repetitions based on lexical and semantic similarity, achieving AUC scores of more than 99.7% in both languages.

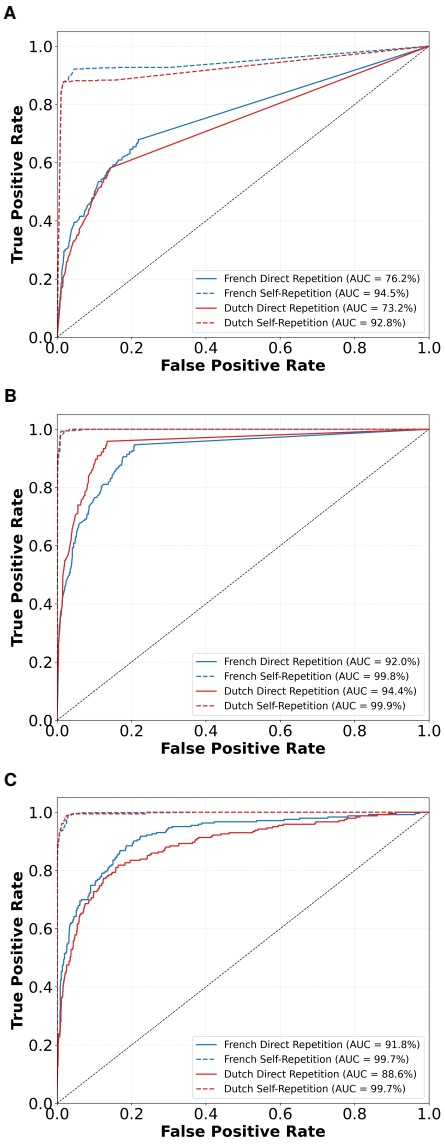


Figure 1: ROC curves and corresponding AUC values for syntactic (A), lexical (B), and semantic (C) cosine similarity models, comparing French and Dutch datasets across direct and self-repetition types

In the following, we will illustrate the observed differences in the distributions of the linguistic measures for repetitive vs. non-repetitive segment pairs in both phenomena in the two languages. Figure 2 shows the distribution for candidates for direct repetition, and Figure 3 for self-repetition. The thresholds that achieved the best precision-recall combination are indicated as reference lines on the box plots.

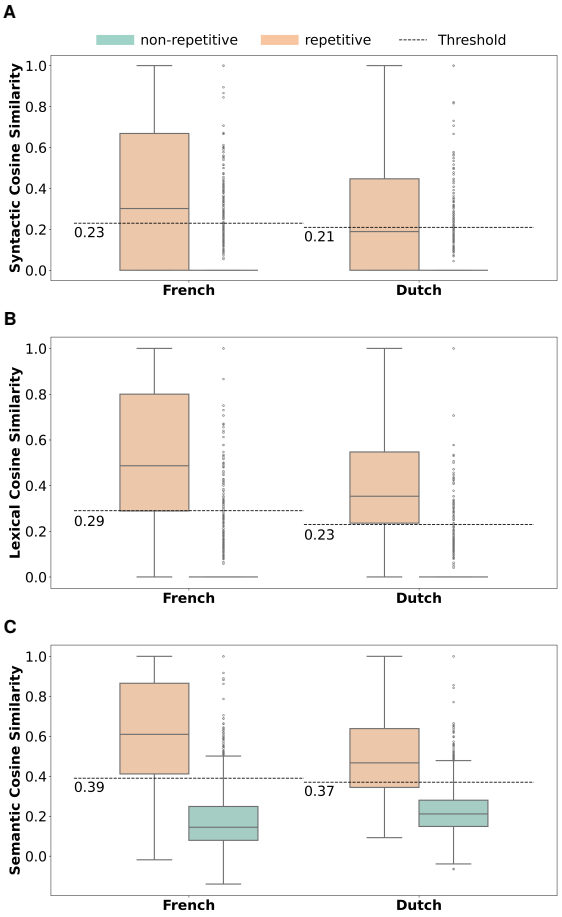


Figure 2: Distributions of syntactic (A), lexical (B), and semantic (C) cosine similarity measures for direct repetition versus non-repetitive segment pairs in the French and Dutch datasets

3.1.1. Performance of the Models Detecting Direct Repetitions

According to Table 2, the best overall results for detecting direct repetitions are achieved using thresholds based on lexical and semantic cosine similarity, yielding recall rates of 76.1% and 75.2% for French and Dutch, respectively. However, the low precision values suggest a high proportion of false positives.

Furthermore, Figure 2 shows that the distribution of the similarity values does not show the ex-

Phenomenon	Model	Language	Threshold	Precision	Recall	F1-score
Direct repetition	Syntactic CosSim	FR	0.23	41.2%	58.0%	48.2%
		DU	0.21	39.1%	47.9%	43.0%
	Lexical CosSim	FR	0.29	59.3%	73.7%	65.7%
		DU	0.23	60.3%	75.2%	66.9%
	Semantic CosSim	FR	0.39	55.9%	76.1%	64.5%
		DU	0.37	52.0%	68.6%	59.2%
Self-repetition	Syntactic CosSim	FR	0.90	61.5%	84.3%	71.1%
		DU	0.88	46.5%	85.0%	60.1%
	Lexical CosSim	FR	0.88	87.9%	88.8%	88.3%
		DU	0.92	86.5%	89.1%	87.8%
	Semantic CosSim	FR	0.88	87.8%	89.0%	88.4%
		DU	0.88	86.8%	87.8%	87.3%

Table 2: Evaluation of repetition detection across phenomena, models, and languages, reporting precision, recall, and F1-score, alongside the optimal Cosine Similarity thresholds. Results are computed over the full dataset

pected pattern (i.e., non-repetitive pairs concentrated in the lower part and repetitive pairs in the higher part of the plot). While non-repetitive pairs are largely concentrated in the lower range of the plots, a significant proportion of outliers appear in the upper range, particularly for syntactic similarity. Moreover, the distribution of repetitive pairs exhibits considerable dispersion. Consequently, a substantial number of repetitive pair values fall below the thresholds and are thus not detected as repetitive. Additionally, the threshold values for direct repetitions are markedly lower than those for self-repetitions, indicating a reduced degree of linguistic overlap between segment pairs.

Lastly, cosine similarity distributions and selected thresholds vary between languages, with consistently lower values for Dutch than for French. This difference is most pronounced in lexical similarity, where the optimal threshold is 0.29 for French and 0.23 for Dutch.

3.1.2. Performance of the Models Detecting Self-Repetitions

The box-plots in Figure 3 illustrate the distribution of similarity measures for self-repetitions versus non-repetitive pairs. As expected, non-repetitive pairs predominantly exhibit low similarity values, whereas repetitive pairs show high values. The thresholds for all measures consistently exceed 0.8, effectively dividing the plots into two distinct areas with relatively few outliers on either side. Moreover, these thresholds remain highly similar across both languages. These observations suggest that self-repetitions are characterized by substantial overlap across all linguistic levels (syntactic, lexical, and semantic).

Nevertheless, differences in distribution are evident across measures. Syntactic similarity plots display greater dispersion in similarity scores, with notably more repetitive outliers in the lower range (0.0-0.6 cosine similarity) and more non-repetitive outliers above the threshold (0.88 or 0.90) compared to lexical and semantic measures. Consequently, the syntactic similarity threshold results in overall lower precision values, particularly for the Dutch data (French: 61.5%, Dutch: 46.5%) in contrast to precision scores between 86.5% and 87.9% for other measures (Table 2). Additionally, cosine similarity scores for non-repetitive segment pairs are generally more concentrated in the lower range (0 - 0.2) for Dutch than for French, except for semantic cosine similarity scores.

Recall scores are high for all thresholds, particularly for lexical and semantic similarity, ranging between 84.3% and 89.1%, with the highest values in lexical and semantic cosine similarities. These results indicate that high lexical and semantic similarity serve as robust cues for distinguishing self-repetitions from non-repetitive segment pairs by the same speaker.

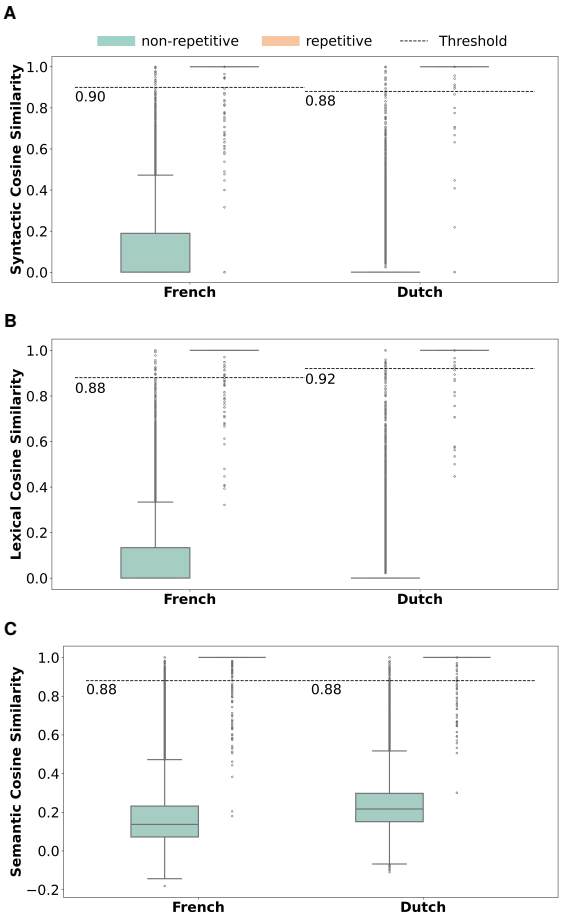


Figure 3: Distributions of syntactic (A), lexical (B), and semantic (C) cosine similarity measures for self-repetition versus non-repetitive segment pairs in the French and Dutch datasets

3.2. Comparison of the Performance of Cosine Similarity and BERT Models on Test Set

While the previous sections reported the results of cosine similarity models applied to the entire dataset, here we compare their performance with BERT models on BERT's test set (8 of 29 speakers).

Overall, BERT models achieve the highest F1-scores for direct repetition in both languages (Table 3), with the French model reaching 81.5% and the Dutch one 73.2% In contrast, for self-repetition (Table 4), lexical and semantic cosine similarity models perform comparably or slightly better than BERT in French, while BERT achieves the highest scores in Dutch (F1-score = 94.0%). Across both phenomena, syntactic cosine similarity models consistently show lower performance. These results suggest that BERT models are particularly effective for detecting direct repetition, whereas lexical and semantic similarity provide competitive performance for self-repetition, especially in French.

Model	Lang	Prec.	Rec.	F1
Syntactic CosSim	FR	45.2%	54.6%	49.4%
	DU	39.6%	40.4%	40.0%
Lexical CosSim	FR	64.7%	71.6%	68.0%
	DU	64.0%	73.7%	68.5%
Semantic CosSim	FR	64.6%	75.0%	69.4%
	DU	49.6%	68.7%	57.6%
BERT	FR	**78.6%**	**84.6%**	**81.5%**
	DU	**66.1%**	**82.1%**	**73.2%**

Table 3: Performance of models on direct repetition detection across languages, reporting precision, recall, and F1-score

Model	Lang	Prec.	Rec.	F1
Syntactic CosSim	FR	48.9%	75.2%	59.3%
	DU	46.4%	75.2%	57.4%
Lexical CosSim	FR	81.6%	**88.3%**	**84.8%**
	DU	84.9%	86.5%	85.7%
Semantic CosSim	FR	**85.7%**	80.8%	83.2%
	DU	85.7%	80.8%	83.2%
BERT	FR	78.0%	80.7%	79.3%
	DU	**92.8%**	**95.2%**	**94.0%**

Table 4: Performance of models on self-repetition detection across languages, reporting precision, recall, and F1-score

4. Discussion

Extending the approach proposed by Fusaroli et al. 2023, this study computed cosine similarity across syntactic, lexical, and semantic representations to detect repetition patterns in autistic children's speech. Overall, the approach proved effective, particularly for self-repetitions.

In the case of *direct repetitions*, models detected a substantial portion of repetitions (recall around 75% or higher) using lexical and semantic similarity measures; however, precision remained limited due to numerous false positives. These findings suggest that predictions for direct repetitions should be interpreted with caution. The limitation likely arises from the annotation protocol, which labels segments as direct repetitions even when they share only a single content word (e.g., Adult: "Do you want a banana?" Autistic Child: "I like bananas"). Because such overlap represents only a small portion of the overall vector representation (especially in longer segments), vector-based similarity measures may not adequately capture these cases. In contrast, BERT models achieved more

balanced performance, with precision above 65% and recall above 80%, suggesting that contextualized representations are better suited for detecting direct repetitions.

Differently, in detecting *self-repetitions*, lexical and semantic similarity-based models performed consistently well across languages and remained competitive with BERT models. However, BERT models showed notable variability, with an F1-score of 94.0% for Dutch compared to 79.3% for French. This discrepancy likely reflects differences in generalization from training to test data rather than data size alone, as both models were trained on substantial datasets. The detection of self-repetitions requires capturing deeper linguistic relationships (e.g., dependency structures; cf., annotation protocol), which may not have been equally well learned by the models across languages.

Across all analyses, lexical and semantic similarity emerged as the most reliable indicators of repetition, yielding high precision and recall scores. In contrast, syntactic similarity consistently showed lower performance, suggesting that syntactic structure in spontaneous speech is highly variable and difficult to capture with surface-level representations. More advanced syntactic modeling may therefore be required to improve performance.

This study highlights the potential of similarity-based approaches for analyzing spontaneous speech in naturalistic contexts. Future work should extend this framework to a broader range of languages and age groups to explore how repetition patterns vary across different linguistic and developmental contexts. A more systematic investigation of cross-linguistic differences (both linguistic and methodological) could further clarify performance variation.

For instance, similarity distributions and thresholds differed between French and Dutch, with higher values observed for French in direct repetitions, whereas self-repetition results were largely comparable. This pattern may reflect differences in interaction styles or overall verbal output, but could also be influenced by language-specific properties or limitations of the NLP tools used. Indeed, French- and Dutch-based spaCy and SentenceBERT models are trained on relatively limited datasets compared to English resources and are optimized for written language, which may reduce their effectiveness on spontaneous child speech. Future research should therefore compare alternative models and embeddings, including those trained on spoken or child-directed data.

While our study demonstrates the effectiveness of cosine similarity-based models for detecting self-repetitions, the challenges in detecting direct repetitions highlight the need for refined methods. For instance, lemma-based rule systems or adaptive thresholding techniques could improve detection of direct repetitions. The observed differences between French and Dutch further suggest that both linguistic structure and NLP model limitations influence performance, underscoring the need for additional cross-linguistic exploration. Future research should also evaluate multilingual or fine-tuned models to enhance repetition detection across languages and spontaneous speech contexts.

We encourage interested researchers to test our models on their conversational data while considering the potential limitations. To facilitate this, the models presented in this paper are publicly available at this GitHub repository. In the similarity-based models, users can select linguistic levels for comparison (syntactic, lexical, and semantic) and adjust cosine similarity thresholds. They are not restricted to the thresholds presented in this paper but may experiment with values within an acceptable range. Users can also test the trained BERT models on their own (French or Dutch language) data.

Finally, a key limitation of this study is the absence of a widely accepted definition of echolalia that allows for fully automated detection. Our annotation protocol attempts to address this issue using simplified linguistic criteria (e.g., comparing lemmas, POS tags, and dependency structures between linguistic units). However, this simplification introduces constraints. For example, evaluating similarity at the segment level may obscure word-level repetition patterns, and some detected cases may not align with traditional definitions of echolalia. Accordingly, the proposed models should be understood as an initial filtering step rather than a definitive classification tool. Establishing clearer and more consistent definitions of echolalia will be essential for improving future detection methods.

5. Ethics Statement

I testify on behalf of all co-authors that the present article was submitted following ethical principles in publishing. All authors declare no conflict of interest and agree that this research presents an accurate account of the work performed. All data presented are accurate, and methodologies are detailed enough to permit others to replicate the study. We share all code used to produce the work, including gold standard annotation protocol, inter-rater agreement calculation, model development and evaluation, and creation of the tables and graphs in the paper. In our code repositories, we provide instructions to interested users on how to apply our code to their own datasets.

However, raw and sensitive data such as speech transcriptions and speaker identifiers cannot be shared to protect the privacy and security of the participants. Ethical approval was obtained from the Ethics Committee of Erasme Hospital on 28 March 2023 (CCB B4062023000074). Written informed parental consent and minor assent were obtained from all participants prior to enrollment in the BeLAS study.

6. Limitations

This study has several limitations that should be acknowledged to contextualize its findings and inform future research.

First, a significant limitation lies in the lack of a universally accepted definition of echolalia. To facilitate detection, we employed simplified linguistic criteria designed for potential automation. While effective in some cases, this approach led to the identification of certain segments that do not qualify as true echolalic instances (e.g., single-word vocatives, such as names or calls, repeated during the recording). Conversely, it also failed to capture echolalic phrases that did not align with the adopted definition, such as repetitions involving the same word used in different syntactic structures. The trade-off between simplicity and comprehensiveness highlights the need for more precise definitions of echolalia. Establishing clearer criteria would improve the reliability and validity of automated detection methods, ensuring better alignment with the nuanced patterns of echolalic speech.

Second, technical challenges associated with pre-trained NLP models must be addressed. The tools used in this study, including BERT, BERT-SentenceTransformers, and spaCy, exhibited variable performance across the two analyzed languages. These models are typically optimized for formal written text and are not designed to account for the unique characteristics of spontaneous children's speech. As such, they may struggle to process features such as informal grammar, incomplete sentences, or age-specific vocabulary. Further fine-tuning NLP models specifically for spontaneous speech data could significantly enhance the accuracy and reliability of repetition detection in this domain. Moreover, the quality of these models varies by language, with NLP algorithms for French and Dutch generally being less robust than their English counterparts due to more limited training data. Future research could benefit from employing more advanced or domain-specific NLP models to mitigate these limitations.

Third, the transcription protocol used in this study introduces additional constraints. Specifically, a new linguistic unit was defined when there was a pause of one second or longer in the child's speech. While necessary for standardization, this approach may have inadvertently excluded pairs of self-repetitions with different syntactic structures simply because they were followed by another linguistic unit. This limitation underscores the need for more flexible transcription criteria that account for the temporal dynamics of naturalistic speech or for a more precise definition of the phrase unit to be considered during comparisons.

Fourth, our analysis revealed potential language-specific variability in repetition patterns and model performance. For instance, thresholds for detecting direct repetitions were consistently higher in French than in Dutch. This variability raises questions about the generalizability of the established thresholds to other languages. Additionally, the lack of validation on independent datasets limits the broader applicability of our models, particularly for detecting direct repetitions. Future studies should test these models across diverse linguistic contexts to refine their utility and generalizability.

Fifth, limitations in the syntactic representations used in this study must also be noted. For syntactic vectors, spaCy was used to extract POS tags, which were grouped into n-grams (n=*2*). While this approach facilitated uniform vector structures, it introduced potential biases when the linguistic segment contained fewer tokens than the selected n, resulting in less informative representations. Additionally, the inclusion of function words may have had minimal influence on similarity measures. Further exploration of alternative vectorization strategies, such as experimenting with different values of n, is warranted to address these concerns.

Sixth, we explored fine-tuning BERT models for repetition detection using our gold standard annotated data. While this yielded promising results for direct repetition, performance for self-repetitions was less reliable, and precision differed significantly between languages. Future research should investigate these differences, using larger training sets and alternative (e.g., multilingual) classification models.

Despite these limitations, the methodology and findings presented in this study provide a valuable foundation for advancing the automated detection of direct and self-repetitions. Future research should aim to refine these methods and extend their application to a wider range of languages, age groups, and conversational contexts.

7. Acknowledgements

This work was supported by the Excellence of Science grant (FNRS and FWO) for the Belgian Language in Autism Study (BeLAS). We thank all the

families and children who participated in the study. We are grateful to our colleagues in the BeLAS consortium for their valuable contributions, and to the research assistants for their help with coding and data management.

8. References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5 edition. American Psychiatric Publishing, Arlington, VA.

S. Amiriparian, A. Baird, S. Julka, A. Alcorn, S. Ottl, S. Petrović, E. Ainger, N. Cummins, and B. Schuller. 2018. Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks. In *Proceedings of Interspeech 2018*, pages 2334–2338.

Nancy Bayley. 2006. *Bayley Scales of Infant and Toddler Development*, third edition. Harcourt Assessment, San Antonio.

B. Bigi, R. Bertrand, and M. Guardiola. 2014. Automatic detection of other-repetition occurrences: Application to french conversational speech. In *Proceedings of Speech Prosody 2014*.

P. Boersma and D. Weenink. 2025. Praat: doing phonetics by computer [computer program]. Version 6.4.26, retrieved 8 January 2025 from http://www.praat.org/.

P. Delobelle, T. Winters, and B. Berendt. 2020. Robbert: a dutch roberta-based language model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Lloyd M. Dunn and Douglas M. Dunn. 2019. *Échelle de vocabulaire en images Peabody, Cinquième édition (EVIP-5)*. Pearson, Paris.

R. Fusaroli, E. Weed, R. Rocca, D. Fein, and L. Naigles. 2023. Repeat after me? both children with and without autism commonly align their language with that of their caregivers. *Cognitive Science*, 47(11):e13369.

M. Honnibal and I. Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia. 2021. An open-source voice type classifier for child-centered daylong recordings.

C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, and S. Bishop. 2012. *Autism diagnostic observation schedule, second edition (ADOS-2)*. Western Psychological Services.

R. J. Luyster, E. Zane, and L. Wisman Weil. 2022. Conventions for unconventional language: Revisiting a framework for spoken language features in autism. *Autism & Developmental Language Impairments*, 7:23969415221105472.

P. Maes, C. La Valle, and H. Tager-Flusberg. 2024. Frequency and characteristics of echoes and self-repetitions in minimally verbal and verbally fluent autistic individuals. *Autism & Developmental Language Impairments*, 9:23969415241262207.

M. K. Marom, A. Gilboa, and E. Bodner. 2018. Musical features and interactional functions of echolalia in children with autism within the music therapy dyad. *Nordic Journal of Music Therapy*, 27(3):175–196.

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot. 2020. Camembert: A tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

T. C. McFayden, S. M. Kennison, and J. M. Bowers. 2022. Echolalia from a transdiagnostic perspective. *Autism & Developmental Language Impairments*, 7:23969415221140464.

OpenAI. 2025. Chatgpt.

Ortho Édition. 2009. *Évaluation du langage oral de l'enfant de 2 ans 3 mois à 6 ans 3 mois*. Ortho Édition, Isbergues.

E. Pascual, A. Dornelas, and T. Oakley. 2017. When "goal!"means 'soccer': Verbatim fictive speech as communicative strategy by children with autism and two control groups. *Pragmatics & Cognition*, 24(3):315–345.

N. Reimers and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

N. Reimers and I. Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

S. Ryan, J. Roberts, and W. Beamish. 2024. Echolalia in autism: A scoping review. *International Journal of Disability, Development and Education*, 71(5):831–846.

J. Schaeffer, M. Abd El-Raziq, E. Castroviejo, S. Durrleman, S. Ferré, I. Grama, P. Hendriks, M. Kissine, M. Manenti, T. Marinis, N. Meir, R. Novogrodsky, A. Perovic, F. Panzeri, S. Silleresi, N. Sukenik, A. Vicente, R. Zebib, P. Prévost, and L. Tuller. 2023. Language in autism: Domains, profiles and co-occurring conditions. *Journal of Neural Transmission*, 130(3):433–457.

Liesbeth Schlichting. 2005. *Peabody Picture Vocabulary Test–III–NL: Nederlandse versie. Handleiding*. Harcourt Test Publishers, Amsterdam.

Eleanor Semel, Elisabeth H. Wiig, and Wayne A. Secord. 2020. *Clinical Evaluation of Language Fundamentals Preschool–3: Dutch Version*. Pearson Assessment, Amsterdam.

Peter J. Tellegen and Jacobus A. Laros. 2017. *SON-R 2–8: Snijders-Oomen Niet-Verbale Intelligentietest*. Hogrefe, Göttingen.

J. P. H. van Santen, R. W. Sproat, and A. P. Hill. 2013. Quantifying repetitive speech in autism spectrum disorders and language impairment. *Autism Research*, 6(5):372–383.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

F. Xie, E. Pascual, and T. Oakley. 2023. Functional echolalia in autism speech: Verbal formulae and repeated prior utterances as communicative and cognitive strategies. *Frontiers in Psychology*, 14.

Author Index