LREC 2026

The 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026

Workshop Proceedings

Editors:

Piotr Bański
Dawn Knight
Marc Kupietz
Andreas Witt
Alina Wróblewska

May 11, 2026

Proceedings of the 12th Workshop on
Challenges in the Management of Large Corpora (CMLC-12)

ISBN 978-2-493814-67-8

Preface

The twelfth meeting of the workshop on the Challenges in the Management of Large Corpora has brought us back to where it all started: a welcoming embrace of the LREC conference series. As in the previous CMLC meetings, we have decided to explore common areas of interest across a range of issues in language resource management, corpus linguistics, natural language processing, natural language generation, and data science.

Large textual datasets require careful design, collection, cleaning, encoding, annotation, storage, retrieval, and curation to be of use for a wide range of research questions and to users across a number of disciplines. A growing number of national and other very large corpora are being made available, many historical archives are being digitised, numerous publishing houses are opening their textual assets for text mining, and many billions of words can be quickly sourced from the web and online social media.

A mixed blessing of our times is that much of those texts, in mono- and multi-lingual arrangements, can now be created automatically by exploiting Large Language Models at various scales. That, on the one hand, makes it possible to inflate the amounts of data where normally data would be scarce: in under-resourced languages or language varieties, in specific genres or for intricate and rarely attested constructions. On the other hand, such procedures immediately raise concerns regarding the authenticity and quality of such data, casting doubt on the possibility of adequately (truthfully, verifiably, reproducibly) addressing the kind of research questions that prompted the rapid but tainted increase of the available data volumes in the first place. Similar doubts may be directed at mass creation of secondary and tertiary data ordinarily crucial for linguistic research: apart from potential legal constraints on the use of the original body of human-created data, new questions arise as to the legal status of the derived data, the ways to create (among others) provenance metadata of the derived resources, and the level of trust regarding mass-produced grammatical (and other) annotation layers.

These new as well as more traditional questions lie at the base of the list of topics that management of large corpora (for any currently suitable definition of "large") invokes or at least strongly brushes against.

CMLC has often invited reports on broadly understood national corpus initiatives. Given that it has been a while since the last round, we have decided to host a "What's the news?" session, with some of our veteran presenters as well as colleagues who have not yet introduced their national corpus projects. The scale of the response to our invitation has been a welcome surprise and we are happy to devote part of this volume to (broadly defined) national corpus projects. This is intended as a snapshot of the current state, with a proud look at the history (including developments since some of the projects were last presented at previous CMLCs), and a somewhat wary glance at what the future may bring.

P. Bański, D. Knight, M. Kupietz, A. Witt, A. Wróblewska
April 2026

Organising Committee

Piotr Bański (IDS Mannheim)
Dawn Knight (Cardiff University)
Marc Kupietz (IDS Mannheim)
Andreas Witt (IDS Mannheim)
Alina Wróblewska (ICS PAS, Warsaw)

Programme Committee

Laurence Anthony (Waseda University, Japan)
Vladimír Benko (Slovak Academy of Sciences)
Felix Bildhauer (IDS Mannheim)
Mark Davies (English-Corpora.org)
Nils Diewald (IDS Mannheim)
Kaja Dobrovoljc (University of Ljubljana / Jožef Stefan Institute)
Jarle Ebeling (University of Oslo)
Tomaž Erjavec (Jožef Stefan Institute, Ljubljana)
Andrew Hardie (Lancaster University, UK)
Serge Heiden (ENS de Lyon)
Ulrich Heid (University of Hildesheim)
Nancy Ide (Vassar College / Brandeis University)
Olha Kanishcheva (Heidelberg University)
Gražina Korvel (Vilnius University)
Natalia Kocyba (Samsung Poland)
Michal Křen (Charles University, Prague)
Anna Latusek (ICS PAS, Warsaw)
Paul Rayson (Lancaster University)
Laurent Romary (INRIA)
Thomas Schmidt (University of Duisburg-Essen)
Serge Sharoff (University of Leeds)
Maria Shvedova (Kharkiv Polytechnic Institute / University of Jena)
Irena Spasić (Cardiff University)
Martin Wynne (University of Oxford)

Table of Contents

Section I: Managing Large Corpora

Section II: News from National Corpus Initiatives

Workshop Programme

9:00–9:10 *Welcome and introduction*
Piotr Bański

9:10–9:40 Session A: Short papers
Chair: Dawn Knight

9:10–9:25 *TestiMole-Conversational: A 30-Billion-Word Italian Discussion Board Corpus (1996–2024) for Language Modeling and Sociolinguistic Research*
Matteo Rinaldi, Rossella Varvara and Viviana Patti

9:25–9:40 *IfGPT, a Large Dataset Representing Bulgarian, with the Bulgarian National Corpus as Its Core*
Svetla Peneva Koeva and Ivelina Stoyanova

9:45–10:25 Session B: Flash presentations of posters
Chairs: Alina Wróblewska, Piotr Bański

Hellenic National Corpus: The Current State
Maria Gavriilidou and Nikolaos Sidiropoulos

Corpas Náisiúnta Na Gaeilge 2022-2029: A Project Overview
Mícheál J. Ó Meachair, Úna Bhreathnach, Kevin Scannell, Michal Mechura, Brian Ó Raghallaigh and Gearóid Ó Cleircín

General Regionally Annotated Corpus of Ukrainian: Recent Developments and Future Plans
Maria Shvedova

Recent Developments of the Bulgarian National Corpus
Svetla Peneva Koeva and Ivelina Stoyanova

The British National Corpus 1994 to 2026
Martin Wynne and Megan Bushnell

The Corpus of Contemporary Polish: 2011-2020 Decade and Beyond
Witold Kieraś, Małgorzata Marciniak, Katarzyna Krasnowska-Kieraś and Marcin Woliński

Building the v4 of the Croatian National Corpus
Marko Tadić, Vanja Štefanec and Daša Farkaš

Managing Growth in a National Corpus: The Hungarian National Corpus 3.0 (MNSZ3)
Noémi Ligeti-Nagy, Enikő Héja, Ágnes Bánfi, Flóra Földesi, Bence Sárossy, Boglárka Skrabák, Tamás Váradi and Gábor Prószéky

CoRoLa Version 2.0: Corpus Enrichment and a New Annotation Level
Elena Irimia, Verginica Barbu Mititelu, Radu Ion, Vasile Pais, Maria Mitrofan and Dan Ioan Tufis

The German Medical Text Corpus: Early 2026 Update
Justin Hofenbitzer, Christina Lohr, Frank Meineke, Markus Löffler and Martin Boeker

From Corpus to Community: New NLP Tools for Welsh Language Research and Learning
Dawn Knight and Fernando Alva-Manchego

Swiss-AL: Language Data Platform for Applied Sciences
Julia Krasselt, Philipp Dreesen, Dolores Lemmenmeier-Batinić, Sooyeon Geckeler, Klaus Rothenhäusler and Matthias Fluor

EuReCo, KorAP and DeReKo: Updates on Ingestion and Annotation Pipelines, Backend, Interfaces, Operation, and Corpora
Marc Kupietz, Nils Diewald, Harald Lüngen, Eliza Margaretha Illig, Helge Stallkamp, Uyen-Nhu Tran and Rameela Yaddehige

Merimënga: A Manifest-First Pipeline for Reproducible Albanian Web Corpus Construction
Besim Kabashi and Michael Ruppert

10:40–11:25 Session 3: Poster session
Chair: Andreas Witt

11:30–13:00 Session 4: Long papers
Chair: Marc Kupietz

11:30–12:00 *Pop Lyrics through Time: Challenges in Corpus-Based Modeling of Linguistic and Emotional Dynamics in German Pop Lyrics*
Roman Schneider

12:00–12:30 *The Infrastructure behind Latvian National Corpora Collection*
Roberts Dargis and Baiba Valkovska

12:30–13:00 *Optimized for AI: Curating the Icelandic Gigaword Corpus for Stable LLM Training*
Jón Friðrik Daðason and Steinþór Steingrímsson

TestiMole-Conversational: A 30-Billion-Word Italian Discussion Board Corpus (1996–2024) for Language Modeling and Sociolinguistic Research

Matteo Rinaldi*, Rossella Varvara*, Viviana Patti*
*Dipartimento di Informatica, University of Turin, Italy
Corso Svizzera 185, 10149 Torino, Italy
matteo.rinaldi@unito.it, rossella.varvara@unito.it, viviana.patti@unito.it

Abstract

We present TestiMole-Conversational a massive collection of discussion boards messages in the Italian language. The large size of the corpus, almost 30B word-tokens (1996–2024), brings challenges in the processing and curation of the resource, but it renders it an ideal dataset for native Italian Large Language Models' pre-training. Furthermore, discussion boards' messages are a relevant resource for linguistic as well as sociological analysis. The corpus captures a rich variety of computer-mediated communication, offering insights into informal written Italian, discourse dynamics, and online social interaction in a wide time span. Beyond its relevance for NLP applications such as language modelling, domain adaptation, and conversational analysis, it also support investigations of language variation and social phenomena in digital communication.

Keywords: Italian language corpus, pre-training data, discussion forums, diachronic corpus

1. Introduction

Over the past three decades, a new form of written communication has emerged due to the diffusion of digital communication networks among the general public. This constituted a revolutionary event in the history of written language, as the digital medium began to be massively used as a form of communication meant for ordinary, spontaneous and interpersonal conversations, a domain that was previously addressed mainly by the oral form of communication. Researchers have coined different terms to refer to this variety, e.g., *computer-mediated communication* (CMC), *netspeak*, *online communication* (for the Italian language, *e-taliano*, *italiano digitato* or *italiano neomediale*, Pistolesi, 2018). This type of language is at the crossroad between oral and written language: despite the written medium, it has many features of informal oral communications; moreover, it has became a pervasive form of written communication even for less scholarized people, a social group that previously had orality as a primary form of communication (Antonelli, 2016).

However, although there may be common features to all instances of CMC[1], different varieties of this type of communication can be recognized. For instance, the language used by a journalist in a blog post will be different from text messages sent through an instant messaging app among family members. In this landscape, a specific type of communication can be identified in the web-based forums and newsgroups. Among different platforms users have been able to communicate in *discussion boards* in asynchronous way, exchanging ideas on specific topics.

In this work, we present a corpus of messages exchanged within the online Italian-speaking community, gathering data from two technologies used by the general public for online discussion: Usenet's newsgroups and several independently hosted forums[2]. The corpus comprehends 470 millions messages among forums and 90 millions Usenet, for a total of 30 billions word tokens.

The purpose of gathering such a large amount of this specific type of data is threefold. First, the creation of this corpus allows the data-driven linguistic analysis of the specific language variety of discussion boards, a dialogical type of CMC for which large scale studies are lacking in Italian. The diachronic nature of the corpus, where each post is annotated with the exact time of writing, provides the possibility to perform precise analysis regarding the evolution and change of the online language across time, both at the lexical and grammatical levels. It also gives the opportunity to capture specific orthographic forms that emerged and then declined in digital communication, such as SMS-style abbreviations or specific emoticon's styles.

Second, online resources are fragile and susceptible to the passing of time: resources needed to keep the pages online, such as servers or web-

[1] We will use this term as a hypernym for denoting every type of language written with digital tools, not only computers, but also smartphones and tablets.

[2] Unless otherwise specified, the cover term "discussion boards" will be employed to refer to both technological configurations, considered their similarity for the purposes of our work.

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 1–11. Palma, Mallorca. May 11, 2026.

hosting, may cease to be maintained. Even if it is hard or impossible to give an estimate about how many online forums have existed since the advent of the web and are now no longer reachable, it is reasonable to expect that a large number of discussion boards are nowadays forever lost, together with all the messages exchanged by their users. This is the reason why archiving this material is a way to ensure its persistence throughout time, avoiding the risk of losing such a peculiar linguistic resource, as well as the large amount of information contained in the posts.

Lastly, given the large size of the collected resource, it may constitute a viable option also as a pre-training dataset for Large Language Models (LLMs), whose training needs impressive amounts of data. The current techniques for training large language models, whether for generative or classification purposes, require a massive amount of data to reach an acceptable level of performance. This poses concern when the goal is to train language models capable of correctly serving the needs of language communities other than English. In the case of Italian, for example, web-derived data that could be used as training data are several orders of magnitude scarcer than the material available for English[3]. A notable exception is the TWITA social media collection (Basile et al., 2018), a ten-year collection of Twitter posts in Italian (2012–2023), that has been exploited for the learning phase of AlBERTo, a BERT language understanding model for the Italian language (Polignano et al., 2019). Discussion boards' messages are an invaluable resource for training large language models: they provide large amount of written texts, originally written in Italian, usually in a direct, informal, and dialogical style that can be useful for developing systems designed to carry out conversations with users, such as chatbots. Moreover, their messages can be considered as "information goldmine", especially when the topics concern highly technical information that is often hard to find in other resources. This information can be employed in LLM training in order to improve their knowledge on more niche themes. Exposing models to discussion boards' messages may improve problem-solving capabilities, given that a significant portion of the messages exchanged on discussion boards involves users asking for and receiving help to solve practical issues. Moreover, the concept of moderation and the way in which users and moderators handle heated discussion can provide the model with hints of how to correctly recognize and handle potential offensive, provocative, or belligerent tones. For these reasons, we believe that a resource such as TestiMole-Conversational can add a significant value to the dataset used to train models for the Italian language.

In this paper, after a review of the related literature, we describe the corpus collection procedure, designed to retrieve a large amount of clean data.We present some of the challenges in corpus creation, such as clean extraction of data and their anonymization, as well as corpus statistics.

The JSONL files of the dataset are currently available at the following address https://huggingface.co/datasets/mrinaldi/TestiMole.

2. Related Work

The potentiality of gathering large text corpora from discussion boards was already explored more than thirty years ago. (Lund and Burgess, 1996; Burgess and Livesay, 1998), notably, compiled the HAL Corpus by collecting 131 million words from Usenet over the course of February 1995. The HAL Corpus was used to train a model encoding semantic and grammatical meaning by transducing lexical co-occurrence. At the time, Usenet was valued as a reliable source for building cognitive models, owing to its conversational nature and the breadth of topics it covered. The HAL corpus, later extended to 320 million words, was also used to create a large database of proper names (Conley et al., 1999). Corpora derived from Usenet were also used in at least two German language projects: the ELWIS corpus, compiled between 1992 and 1993, and employed to investigate language use in newsgroups (Feldweg et al., 1995); the DeReKo project, the German Reference Corpus (Schröck and Lüngen, 2015), into which an annotated version of the ELWIS corpus was integrated.

Other examples for the English language include the reduced redundancy Usenet Corpus (Shaoul and Westbury, 2013), collected from 2005 to 2011 and containing more than seven billion words, the "Usenet as a text corpus", which comprehends 53,245 articles (Mahoney, 2000), and the corpus collected by Hoffmann (2007), constituted by 773,772 messages from Usenet newsgroups.

As for Italian, the group coordinated by Manuel Barbera at the University of Turin developed the NUNC corpora, the largest collection of material drawn from Italian Usenet prior to the current work (127M tokens, Barbera, 2013). Most notably, NUNC corpora are annotated and also available for other languages (Barbera and Marello, 2011) such as Spanish and French (Barbera et al., 2011).

General web corpora contain, among others, texts from forum discussions as well, but not all resources provide metadata that allow the users to

[3]For example, in the Common Crawl dataset Italian accounts only for 2% of the data, in comparison of 44% of English. See https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

identify forum discussions. The most widely known web corpora include, but are not limited to, the WaCky corpora (Baroni et al., 2009), the COW corpora (Schäfer et al., 2012), the SketchEngine's TenTen family (Jakubíček et al., 2013), and the Aranea corpora (Benko, 2014). It is also worth mentioning that the collection of web corpora gained huge attention starting from the late 2000s, as shown by the intense research activity of the ACL special interest group on the Web as Corpus (SIG-WaC).

To date, TestiMole-Conversational is the largest Italian corpus of Usenet and Forum discussions; it can be successfully used for the development and improvement of NLP tools for the Italian language, as well as enabling unprecedented analysis of the language adopted by Italian speakers in the Web over almost thirty years.

Discussion board conversations have also been used to support sociological and psychological analyses of online communication behaviour. One example is the phenomenon of "trolling", whereby users engage in explicit or covert aggression toward others with the deliberate aim of provoking an emotional response in the victim (Hardaker, 2010).

The selective and topic-specific nature of online discussion boards, particularly forums, makes these platforms fertile grounds for marginalized and extremist groups alike. A notable example is the Italian "Forum dei brutti" ('Forum of the ugly'), which serves as the primary online community for the Italians "Incel" subculture. This forum has been extensively studied due to its pronounced misogynistic stance, examined both from sociological perspectives (Cava and Pasciuto, 2023) and within the context of hate speech detection research in computational linguistics (Gajo et al., 2023; Gemelli and Minnema, 2024).

3. The TestiMole-Conversational Resource

3.1. Discussion Boards

The text sources of the TestiMole-Conversational corpus are two types of discussion boards. Discussion boards are platforms where users can exchange messages on specific topics. Messages, called "posts", are organized in "threads" that refer to a very specific topic of discussion, usually identified by the title of the discussion. Compared with instant messaging platforms, such as IRC (Internal Relay Chat), discussion boards are designed for asynchronous communication, where the same discussion can be continued for an indefinite amount of time. The topics allowed on a specific discussion board depend on the rules and characteristics of the given board: posts are organized in specific arbor-like hierarchies limiting the scope of the topics that are appropriate to write in a specific board; forums, on the other hand, are dependent on the decisions of the administrators so, depending on the platform, they can be more or less specific about the topics that intend to represent. In general, it is possible to open a limited number of discussions related to topics different from the scope of the board, and this situation is indicated by the "off-topic" label, often shortened as "OT". Forums are usually moderated, that is, a restricted group of people is given the power to terminate discussions that are leading to personal attacks, as well as to invite users to avoid exacerbating tones. Moderators can also issue temporary or permanent bans to users who violate forum rules. Cursing, especially on larger and more important forums, is usually forbidden. Note, however, that moderation on Usenet was not generally applied, and texts sent by users were not subject to any kind of censorship or restriction.

Compared to social networks, discussion boards are more focused on restricted topics. Such texts are often highly technical, providing a valuable source of hard-to-find information derived from users personal or professional experience. Discussion board messages, indeed, contain detailed and very specific information written by people who are often passionate and very knowledgeable about what they are discussing. Nonetheless, alongside truthful and useful data, users may share blatant misinformation, personal opinions, or errors, with or without being aware of it.

3.1.1. Usenet

Usenet was the first widely used platform for discussion boards on the Internet. Its inception dates back to 1979, when the Network News Transport Protocol (NNTP) enabled the exchange of messages between servers following a hierarchical organization of newsgroups. Articles propagate across servers by creating copies, with distribution managed through relationships between nodes. Unlike centralized forums hosted on single servers, Usenet operates as a distributed system where independent servers exchange articles through bilateral agreements, with no central authority controlling content distribution beyond community established procedures for creating new groups. The diffusion of Usenet in Italy[4] started significantly later than in other countries. While isolated discussion groups existed earlier, the structured Italian hierar-

[4]The material to reconstruct the history of Italian Usenet was obtained from the official pages of the GCN-IT https://www.news.nic.it/ as well as the personal website of one of the Italian Usenet's founders, Maurizio Codogno: https://xmau.com/usenet/

chy *it.** was formally established in 1994-95 through a coordinated effort led by Alessio F. Bragadini and Stefano Suin in the context of the University of Pisa's SerRA Project. The initiative was formalized at the NIR-IT-2 conference in Milan on December 13, 1994, with the goal of creating a national discussion space independent from single institution or provider. Prior to this, the only Italian-language group was *soc.culture.italian*, which mixed international discussions about Italy with domestic content[5]. The new it.* hierarchy officially began operations in January 1995, initially comprising groups like *it.politica*, *it.sport*, *it.spettacolo*, *it.scienza*, and *it.cultura*, designed to reflect traditional media organization. To generate initial traffic, bidirectional gateways were established with major Italian mailing lists, allowing the same content to be accessed both via newsreaders and list subscriptions. International distribution began in March 1995 and by November 1995, the hierarchy had achieved full integration into the worldwide Usenet system. The creation of new newsgroups followed international Usenet standards, requiring 50 interested users or connection to mailing lists with 150 subscribers.The *Gruppo di Coordinamento NEWS-IT* (GCN) was established as a volunteer working group to manage the hierarchy, the *Request For Discussion* (RFD) and *Call For Votes* (CFV) procedures, and user documentation. The GCN decided to establish second-level sub-hierarchies (like *it.comp*, *it.cultura*, *it.hobby*) to thematically organize groups, and to create *it.binari.** for binary files with mandatory moderation to prevent abuse. They also introduced several innovations: groups could be removed if traffic fell below 100 articles in three months, crossposting was limited to a maximum of 10 newsgroups, and binary file attachments were prohibited outside designated groups. Moderation emerged as a critical tool, with both human moderators and experimental "robomoderation" systems that automatically filtered messages based on technical criteria like crosspost count or binary content. Netiquette rules, sometimes referred to in Italian as "galareteo" or "retichetta", were actively promoted through FAQ files[6]. The GCN faced recurring challenges regarding content control, particularly concerning illegal material and copyright infringement, leading to policy discussions with GARR authorities and commercial providers about legal responsibilities. Each newsgroup required a "manifesto" (*charter*) defining its scope and acceptable topics.

[5] An earlier hierarchy ita.* had been created around 1993 by Marco Negri at the University of Milan's Department of Computer Science but failed due to poor coordination between Italian news servers and limited propagation outside the academic network.

[6] https://www.news.nic.it/doc/emily.html

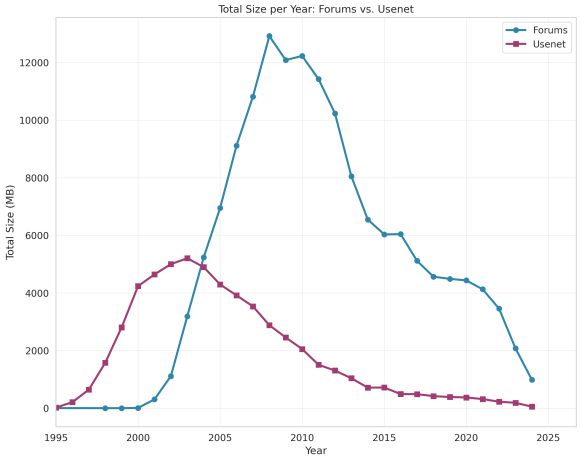


Figure 1: Total corpus size (in MB) per year. Forum overtakes Usenet around 2004.

Newsgroups in Italy were the main form of discussion used in Italy between the late 1990s and the beginning of 00s, alongside with Bullettin Board System (BBS), which were also accessible via a direct dial-up connection. Forums, hosted on private platforms, began to supersede Usenet around 2004, according to our statistics (see fig.1).

3.1.2. Forums

Forums represent an evolution of online discussion spaces after Usenet, offering a more centralized environment for asynchronous communication. Unlike the distributed architecture of Usenet, forums are hosted on dedicated web servers managed by administrators who define both thematic organization and access policies. Each forum is typically structured hierarchically, with sections and subsections grouping discussions according to topic, within which users can post individual messages. Registration is generally required to participate, allowing the system to manage user identities, moderation privileges, and content visibility. Forums usually rely on software frameworks, such as vBullettin, phpBB or Simple Machine Forum, accessible via web-browser instead of a dedicated newsreader software like Usenet. Moderation in forums is usually more formalized than in Usenet: administrators and moderators enforce rules of conduct, prevent abusive or off-topic content, and ensure the overall coherence of discussions. The introduction of moderation tools such as warnings, temporary bans, and message deletion contributed to the development of online communities characterized by hierarchies and social norms.

With their diffusion in the late 1990s and early 2000s, forums became one of the main hubs of online interactions, often forming communities about highly specific interests. Thematic specialization, indeed, was one of the main traits of forums:

distinct platforms emerged for every conceivable topic, from technical discussions to entertainment, forming a form of collective knowledge production. Users' reputations were built solely through the accumulation of contributions, rather than through algorithmic amplification based on engagement metrics, as is the case with current social media. The forum model thus privileged temporal continuity and participation history over immediate visibility. As a form of meritocratic social hierarchy, seniority and posting frequency determined the level of recognition within the community.

A distinctive aspect of forums was their semiotic dimension: users developed and shared linguistic conventions characterized by abbreviations, neologisms, and acronyms (such as "trolling","flame","newbie"/"niubbo").

Despite their subsequent decline with the advent of social networks, forums have persisted as a form of online communication. Their decentralized and topic-oriented architecture constitutes a different paradigm compared to the algorithmically driven and feed-based structure of social networks: forums tend to promote discussions focused on specific contents, while social networks privilege transient engagement.

3.2. Collection Methodology

The material was obtained through a web scraping effort going on from February 2024 to May 2024. The texts more distant in time date back to 1996. To create the data set, several scripts were developed using Python3 and libraries such as BeautifulSoup and Selenium. The scripts were often manually crafted for each resource in a very slow but precise process, even if the overall structure of the code was generally the same. In particular, all the scraped forums were based on a limited number of discussion board platforms, such as vBulletin, phpBB, XenForo, Simple Machines Forum, Invision, and Snitz. Often, it was possible to use the same script designed for a forum based on a specific platform (e.g., phpBB) for a different forum sharing the same platform. However, this was not always the case: forum software usually allows for a great degree of customisation, through administrators' settings or the adoption of plugins, and such differences required different parsing patterns to be configured in each scraping script. Thanks to the flexibility of the BeautifulSoup library, most of the work was limited to identifying the correct HTML identifiers, a task made easier by modern browsers equipped with developer consoles such as Mozilla Firefox. The BeautifulSoup library provides the programmer with helpful *syntactic sugar* that can be used, for example, to find all the elements sharing a specific class, iterate over them, perform additional operations on the identified objects, and so on.

In general, the first step was to identify the URL pattern logic for each topic published in the forum. For example, if the forum discussions have a URL such as `FORUM_BASE_URL/viewtopic.php?t=N`, where `N` uniquely identifies the discussion, it is then possible to cycle through all the discussions from number 0 up to the latest one, which is manually identified by checking the most recently posted message at the time of scraping (although this verification could probably be automated). In some cases, a SEO-friendly string is appended to the URL, for example, `FORUM_BASE_URL/category-name/N/title-of-the-discussion`.

After having identified the URL pattern, the next step was to understand the logic that the discussion board used to handle threads with many messages: all discussion board platforms include a pagination mechanism in order to avoid loading a large number of messages all at once. In general, it was possible to identify a specific HTML identifier for the container of pagination links (e.g., `<div class="pagination">`), and to retrieve the last link in the pagination container.Pagination was one of the trickiest parts of the scraping process, due to the high variability we found across the scraped forums. In some cases, it was possible to obtain a more regular URL schema by using the "printer-friendly" version of the page, when available.

The remaining useful information in the web pages was generally easier to identify. For each discussion, we were interested in retrieving all the messages, and finding the correct rule was straightforward, because most platforms use a unique identifier for each post (e.g., `<div class="post">`, `<li class="message">`, ...), making it possible to iterate over the portions of HTML code pertaining to each post. After having identified the logic for isolating individual posts, it was necessary to locate the identifiers for the required fields, namely the author of the post, the time and date of posting, and the body of the message. Again, manual inspection was used to identify the HTML identifier for each field, although these identifiers were often shared across different forums. Parsing the date was trickier: in the most straightforward cases, a special HTML `<datetime>` tag was present, making it trivial to extract the message timestamp. Conversely, when only the written Italian form of time and date was present (e.g., *Alle 16.30 di Domenica, 14 Dicembre 1997*), more effort was required to reconstruct the machine-readable format of the date (in this case, `1997-12-14T16:30:00`).

Although all the scripts were manually edited and refined for each forum, it is possible to create a ready-to-use pipeline compatible with at least all the forums already scraped, and potentially many more.

This has not yet been accomplished; however, we plan to release a tool ready to use that could be leveraged for other research purposes. Further details about the scripts are reported in Appendix.

From each URL, each post was scraped and added to the dataset as a single row, containing the following metadata:

- "title": //the title of the thread;
- "author": //anonymized ID of the post's author;
- "id": //unique identifier of the thread in the groups' threads;
- "progressive_id": //identifier of the post in the thread; counter starts from 1;
- "timestamp": //the time and data of creation of the post, in ISO-8601 format;
- "newsgroup"/"forum": // the identifier of the specific newsgroup or forum;
- "text": // the body of the message.

In order to provide a pseudonymization of the dataset, each value present in the field "author" was substituted with a progressive number. In this way, the relationship between the authors and the posts was kept, but the specific usernames were hidden.

The computational resources required to run the scraping scripts were minimal: an old dual-core laptop from 2006 was sufficient, and its processing power was not a bottleneck relative to the network speed (domestic ADSL).

The result is a clean dataset that can be used with few preprocessing because the extraction scripts retrieved text data directly in a structured way, without the need to apply further filtering.

While the list of Italian Usenet hierarchies is straightforward to obtain from official sources, for forums we used a search engine to look for potential candidates: the forums selected for scraping were identified using keywords such as `"forum"`, `"viewtopic"`, and `"showthread"`. No strict criteria were adopted for the selection of forums. However, larger forums were preferred, while forums with very few messages (<1,000) or containing only spam were discarded.

3.3. Corpus Statistics

Overall, the TESTIMOLE-CONVERSATIONAL corpus contains almost 30 billions word-tokens[7], with a larger part devoted to forum texts (23 billions vs 7 billions for Usenet data).

Larger amount of data were collected for the central years of the time span considered (between 2003 and 2011, see fig. 2 and 3): pages from the nineties were more rare, probably because no longer maintained. In more recent years, instead,

[7]The exact size of the corpus is 29,592,255,016 tokens.

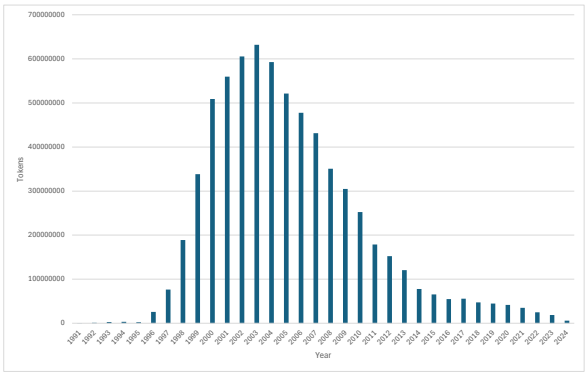


Figure 2: Usenet - Number of tokens per year

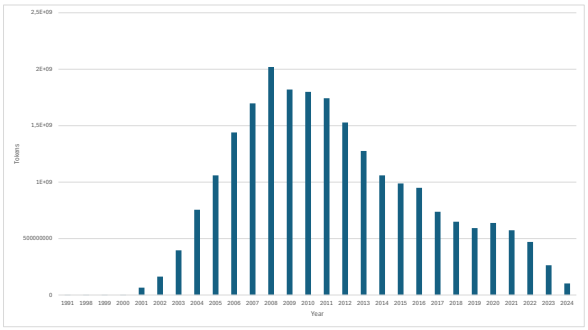


Figure 3: Forums - Number of tokens per year

this type of discussion boards has probably lost success, with other platforms (such as social networks) becoming digital places for discussions. It is interesting to note that the distribution of tokens among years differ for the two subgroups of data: Usenet data reach the peak of tokens in 2003, while for forums data we have the highest number of tokens for the year 2008. It seems indeed that Usenet's popularity decreased earlier than those of other forums (see fig.1).

In terms of number of posts, the Forum corpus contains 468,391,746 posts in 25,280,745 unique threads (on average, 18.5 posts per thread), while the Usenet corpus consists of 89,499,446 posts in 14,521,548 threads (on average, 6 posts per thread). The computation of tokens with a sub-word tokenizer employed for LLM training (Tiktoken BPE tokenizer, model *cl100k_base*) resulted in 20B and 62B tokens for Usenet and Forums, respectively.

The topics covered by the posts are of different nature and can be inferred from the names of newsgroups (fig.4) or forums (fig.5). Among the Usenet section, politics is the first topic covered by the resource, with *it.politica* covering around 6% of the data. Cars and soccer follow as more represented topics. Among forums, the first source is *hwupgrade*, an Italian forum on technology, which represents around 15% of the forum section. The second ranked, *alfemminile*, is a forum devoted to women conversations, with topics that cover pregnancy,

maternity, menstruation, among many others. As for Usenet, politics covers a significant proportion of the dataset, i.e. around 9% of forums data.

The corpus is released together with words frequency lists. Given the diachronic nature of the corpus, frequencies can be used in the socio-linguistic analysis to observe the rising and fall of specific terms in conversation or to identify neologisms. Figure 6 shows the use of six words over the time period of the TestiMole corpus. It shows the rapid growth of use of the neologism *troll*, which was coined during the first years of Usenet groups, and of the neologisms *smartphone* and *streaming*, which appears already on 2001 but gained popularity starting from 2010.

4. The TestiMole Dataset

TestiMole-Conversational is part of a larger dataset originally created in order to provide the academic community with better resources to train different kind of language models employing high-quality native Italian resources, also for long-context training. In this work, we decided to focus on the "conversational" subset as a scientific object on its own, given its relevance for a broad range of studies. It is worth to notice that alongside TestiMole-Conversational, the TestiMole dataset for the Italian language also comprises large amounts of cleaned textual data drawn from public domain books (2GB)[8] as well as open-access academic material (20GB), blogs (15GB) and the collection of several already existing Italian large corpora.

5. Conclusion

In an historical period characterized by the rise of Large Language Models and the consequent quest for large and clean datasets, TestiMole stands out as an important resource for improving the capabilities of natively Italian as well as multilingual LMs to correctly model peculiar elements of Italian language and society, drawing from the rich, diverse and collectively created shared knowledge base compiled by users during thirty years of Computer Mediated Interaction. This material can help models grasp conversational nuances typical of Italian discourse and guide them toward more natural problem-solving paths, potentially richer than those inferred solely from neutral sources. Furthermore, this conversational content can enhance emotion understanding capabilities, leading to improved classification systems for Italian data. Given the proven importance of discussion board analysis for linguistic and sociological research, this resource offers an opportunity to introduce large-scale data analysis into such studies, which are often constrained by limited datasets, thereby reinforcing their experimental validity and unlocking possibilities for novel investigations. In this perspective, future work may include providing linguistic annotation on the collected data, including for example lemmatization, POS tagging and syntactic parsing.

[8] Currently, public domain books were mainly sourced from the LiberLiber's 'Progetto Manuzio' https://liberliber.it/

6. Acknowledgements

The work of V. Patti and M. Rinaldi have been partially supported by the "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme.

7. Limitations

Given the substantial manual work involved in designing appropriate collection strategies from diverse platforms, it was not possible to include every Italian discussion board in the corpus; neither it is possible to quantify the proportion of the collected resource over the total. Further sources could have probably been retrieved, but we believe that the present corpus already represents a wide and representative sample of this variety of CMC language.

It is also important to note that, although moderation was in place for many of the collected resources, we cannot guarantee that the corpus is free from profanity, offensive or aggressive language. Indeed, we did not aim to remove it, since the resource may be used for the study of hate speech as well, even from a diachronic point of view. However, while these registers may be of high interest for specific socio-linguistic research, their usage in language modelling should always be considered in relation to the intended use cases.

Finally, discussion board material may introduce unwanted noise in LLM training compared with data obtained from cleaner sources such as books, encyclopedias or academic papers: users may have introduced, either for amusement or actual misinformation purposes, erroneous or misleading information.

8. Ethical considerations

From an ethical standpoint, the collection of online conversational data raises concerns regarding user privacy and consent, even when such content was publicly accessible at the time of collection. To mitigate these risks, we anonymized all usernames from the corpus. We assume that users followed

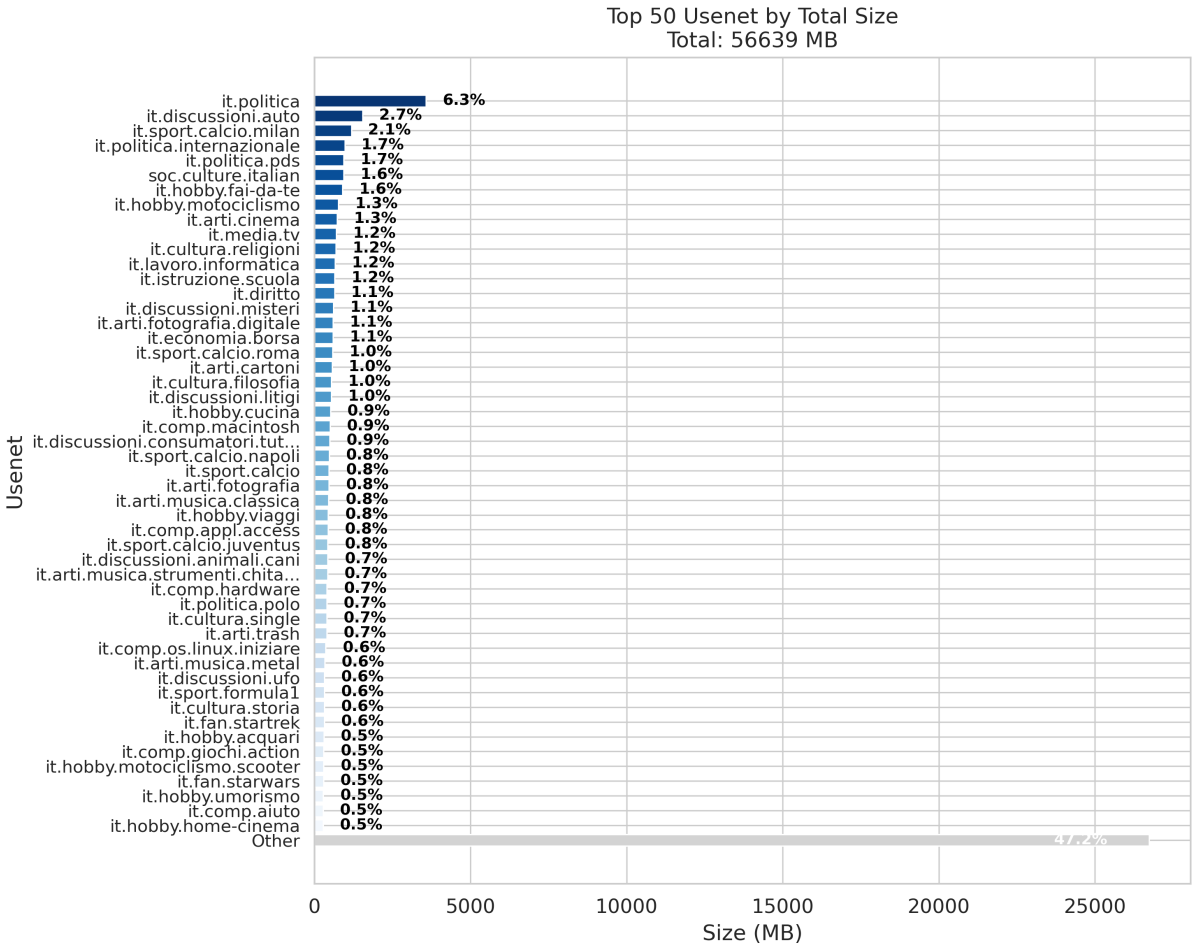


Figure 4: Top 50 newsgroups by total character count (all periods combined).

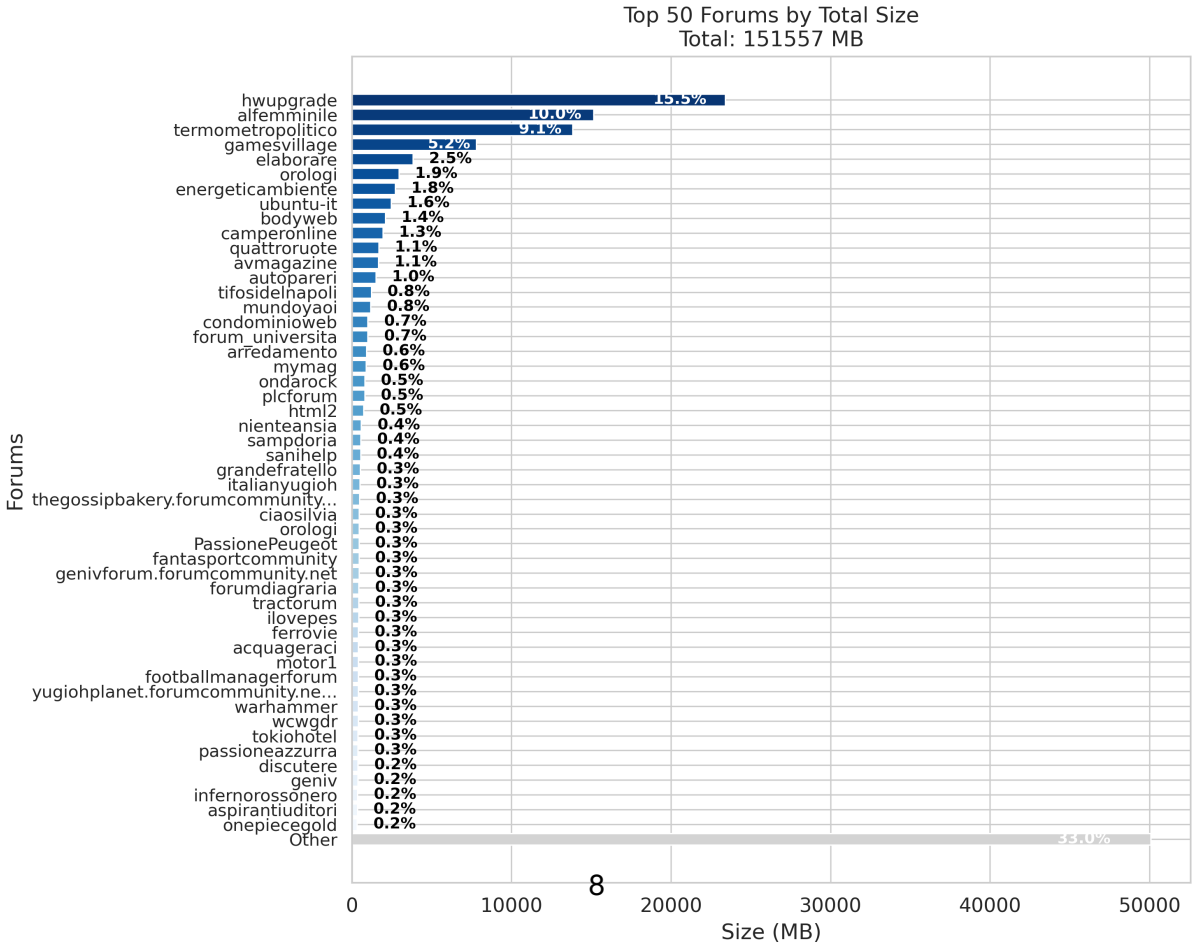


Figure 5: Top 50 forums by total character count (all periods combined).

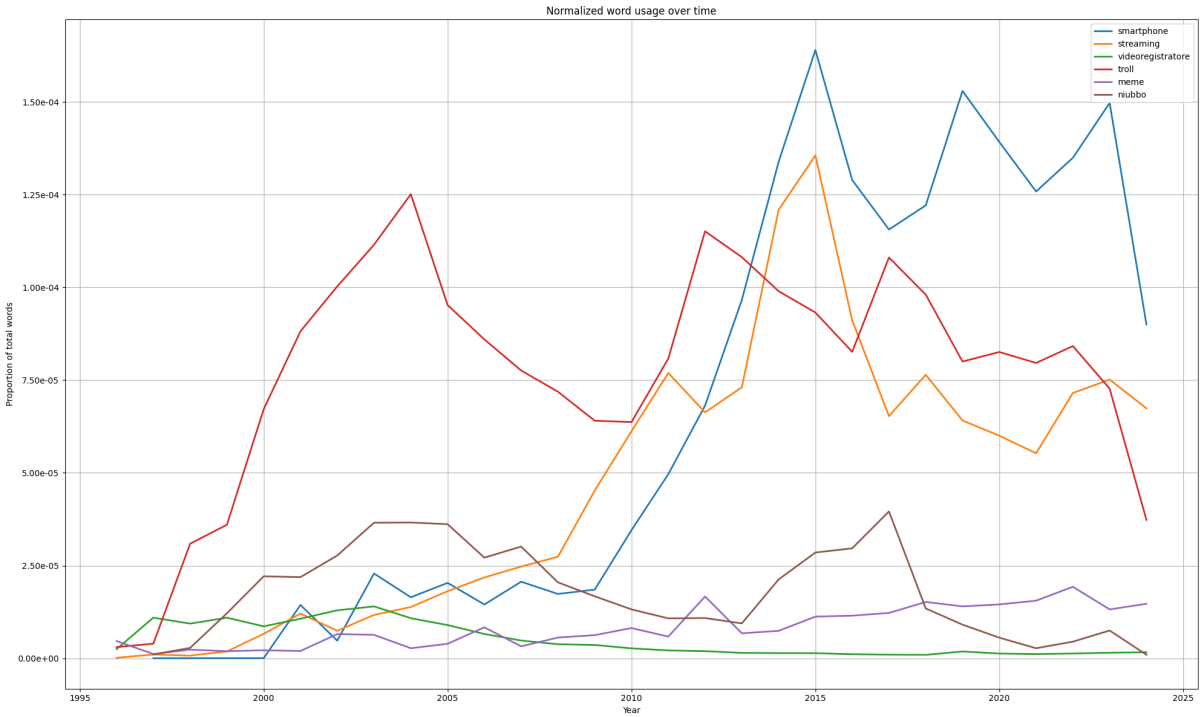


Figure 6: Normalized frequencies of six words across time in the TestiMole-Conversational corpus.

platform guidelines prohibiting the sharing of personal information; however, we acknowledge that inadvertent disclosure of sensitive details may still occur in user-generated content. Researchers using this dataset should be aware of these limitations and exercise caution when analysing or presenting findings that might compromise individual privacy.

According to the current version of the European Digital Services Act (Reg. UE 2022/2065), researchers can download publicly available data from web platforms for research purposes. Indeed, our resource is meant only for research, with no commercial intention. Accordingly, we plan to redistribute the data exclusively to researchers for research purposes, on a case-by-case basis, under the condition that applicants complete a request form including an end-user non-commercial license agreement.

9. Appendix

The script used to web-scrape the data is composed of four main functions:

- The **main** function, given an URL pattern defined by a prefix, a range, and a suffix, calls the `get_url` function for all forum topics.
- The `get_url` function connects to the server and attempts to retrieve the page. In particular, it uses Python's `requests` library to download the URL, logging whether the download was successful or an error occurred. It then passes the downloaded HTML page to the BeautifulSoup parser in order to obtain the object required for subsequent parsing. Any BeautifulSoup errors are also logged.
- The `extract_thread` function, after having identified the discussion title, iterates through all the post HTML blocks (most often, `div` tags) and extracts the relevant information (author, datetime, with optional conversion to ISO format, and message body). If pagination is detected in the thread, the function is called iteratively (with the flag `inside_pagination` set to `True`) for all the thread's pages until all the messages have been parsed. All this information is stored in a list of Python dictionaries: each element of the list corresponds to a post, while the list itself represents the thread. This list is then passed to the `save_post_to_jsonl` function.
- The `save_post_to_jsonl` function iterates over the list and saves each message to a JSONL file, as a single row. Since Python lists are ordered collections, it also reconstructs the order of the posts, which is recorded in the JSONL entry under the `progressive_number` column. An alternative data structure, such as a single JSONL entry per thread containing a list of all the posts, would have been more efficient; however, the less efficient approach was chosen to ensure

better visualisation within the online Hugging Face platform. Conversion between these different storage formats is nevertheless trivial.

10. Bibliographical References

Giuseppe Antonelli. 2016. L'e-taliano tra storia e leggende. In *L'e-taliano. Scriventi e scritture nell'era digitale*, pages 11–28. Franco Cesati Editore.

Emanuele Ferdinando Barbera. 2013. *Una introduzione ai NUNC: storia della creazione di un corpus*, volume Molti occhi sono meglio di uno: saggi di linguistica generale 2008-12, pages 97–114. Qu. A. S. A. R.

Manuel Barbera and Carla Marello. 2011. Tra scritto-parlato, Umgangssprache e comunicazione in rete: i corpora NUNC. In *«Studi di Grammatica Italiana» XXVII (2008, recte 2011) = Per Giovanni Nencioni. Convegno Internazionale di Studi. Pisa - Firenze, 4-5 Maggio 2009*, pages 157– 185. Le Lettere.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 41–46, Turin, Italy. CEUR Workshop Proceedings.

Vladimír Benko. 2014. Aranea: Yet another family of (comparable) web corpora. In *International Conference on Text, Speech, and Dialogue*, pages 247–256. Springer.

Curt Burgess and Kay Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from kučera and francis. *Behavior Research Methods, Instruments, and Computers*, 30:272–277.

Antonia Cava and Fabrizia Pasciuto. 2023. *Misoginia online: un'analisi netnografica sul forum dei brutti*, pages 469–482. Sette città.

Patrick Conley, Curt Burgess, and Dotty Hage. 1999. Large-scale databases of proper names. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 31:215–9.

Helmut Feldweg, Ralf Kibinger, and Christine Thielen. 1995. Zum sprachgebrauch in deutschen news- gruppen. In *Neue Medien. Osn- abrücker Beiträge zur Sprachtheorie*, pages 143–154. Oldenburg: Red. OBST.

Paolo Gajo, Silvia Bernardini, Adriano Ferraresi, and Alberto Barrón-Cedeño. 2023. Hate speech detection in an Italian incel forum using bilingual data for pre-training and fine-tuning. In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 211–218, Venice, Italy. CEUR Workshop Proceedings.

Sara Gemelli and Gosse Minnema. 2024. Manosphrames: exploring an Italian incel community through the lens of NLP and frame semantics. In *Proceedings of the First Workshop on Reference, Framing, and Perspective @ LREC-COLING 2024*, pages 28–39, Torino, Italia. ELRA and ICCL.

Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, 6(2):215–242.

Sebastian Hoffmann. 2007. Processing internet-derived text—creating a corpus of usenet messages. *Literary and Linguistic Computing*, 22(2):151–165.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *7th international corpus linguistics conference CL*, volume 2013, pages 125–127. Valladolid.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Matt Mahoney. 2000. Usenet as a text corpus. Technical report, Florida Tech, CS Dept. Available online at: https://cs.fit.edu/.

Elena Pistolesi. 2018. L'italiano in rete: Usi, varietà e proposte di analisi. *AggiornaMenti*, pages 17—-26.

Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. AlBERTo: Modeling Italian Social Media Language with BERT. *IJCoL [Online]*.

Roland Schäfer, Felix Bildhauer, et al. 2012. Building large corpora from the web using a new efficient tool chain. In *Lrec*, pages 486–493.

Jasmin Schröck and Harald Lüngen. 2015. Building and annotating a corpus of german-language

newsgroups. In *NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop , September 29, 2015 University of Duisburg-Essen, Campus Essen*, pages 17 – 22.

11. Language Resource References

Barbera, Manuel and Colombo, Simona and Marello, Carla. 2011. *NUNC - A Multilanguage Suite of Newsgroups Corpora*. Università degli Studi di Torino. PID http://www.bmanuel.org/projects/ng-HOME.html.

Shaoul, Cyrus and Westbury, Chris. 2013. *A reduced redundancy USENET corpus (2005-2011)*. Edmonton, AB: University of Alberta. PID http://www.psych.ualberta.ca/ westbury-lab/downloads/usenetcorpus.download.html.

IfGPT, a large dataset representing Bulgarian, with the Bulgarian National Corpus as its core

Svetla Koeva, Ivelina Stoyanova
Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
{svetla,iva}@dcl.bas.bg

Abstract

The paper introduces the IfGPT dataset, which integrates several Bulgarian text collections, including the Bulgarian National Corpus, and applies cleaning, deduplication, and LLM-oriented metadata such as personally identifiable information and bias scores. The composition of the IfGPT dataset is presented, along with the unified metadata schema and metadata management in a graph database, enabling efficient querying and document selection for specific tasks. The main contributions are the integration of multiple Bulgarian text collections into a unified dataset, the development of a standardised metadata schema with graph-based organisation, and the provision of efficient metadata querying mechanisms to support LLM development.

1. Introduction

The development and management of large-scale corpora present interconnected technical, linguistic, and management challenges. From a technical perspective, scalable storage and efficient retrieval of text, metadata, and annotation layers are essential (Mohammadi et al., 2025). From a linguistic and NLP perspective, large reference corpora must be diverse and represent a wide range of language use, including low-resource languages, under-represented phenomena, and historical texts.

The **Bulgarian National Corpus** (**BulNC**)[1] is a standard reference corpus designed to reflect the natural distribution of the Bulgarian language across text types, genres, styles, and time periods, ensuring domain coverage and distributional balance (Koeva et al., 2012). Its main purpose is to support linguistic studies focused on the lexical and grammatical features of Bulgarian, dictionary creation, and the exploration of language change. The Bulgarian texts are annotated using the Bulgarian natural language processing pipeline (Koeva et al., 2020), which integrates several tools for different layers of annotation: tokenisation, part-of-speech tagging, lemmatisation, dependency parsing, word sense annotation, lexical relations (synonyms, hypernyms, and similar adjectives), noun phrase identification, and named entity recognition. Linguistic integrity is maintained by deduplicating texts and removing documents with typographical errors, incomplete sentences, or malformed words.

The large volume of available data, including for Bulgarian, has enabled the development of datasets that encompass linguistic and human knowledge about the world for training large language models (LLMs). The dominant approach is to collect as much data as possible, mainly from the web, and then filter it by cleaning and deduplication, as well as by removing content that could degrade model behaviour, such as toxic content, personally identifiable information (PII), near-duplicate documents that may cause memorisation, and very low-quality text that may introduce noise.

The efforts to develop, maintain, expand, and improve the Bulgarian National Corpus are naturally combined with the compilation, cleaning, maintenance, and enhancement of large Bulgarian datasets for pre-training and fine-tuning LLMs. These efforts have resulted in the creation of the large **BulNC-based dataset** within the project **IfGPT: Infrastructure for Fine-tuning Pre-trained Large Language Models**[2] (the **IfGPT dataset**).

The large IfGPT dataset integrates several Bulgarian datasets, including the Bulgarian National Corpus. Like the BulNC, IfGPT contains authentic Bulgarian language data that is cleaned and deduplicated. The IfGPT description adds some LLM-oriented metadata (i.e. PII scores, bias scores) that a pure reference corpus such as BulNC does not include. Annotation in the IfGPT dataset is also modest compared to the BulNC, limited to sentence markup, which relates to the main purpose of its intended use.

In the next section, we briefly present related work on the development of reference corpora and the position of the BulNC within this context. The main part of the paper is devoted to the composition of the IfGPT dataset, its metadata description, and its management through a graph database that provides various access options.

The main contributions of this work are as fol-

[1] https://dcl.bas.bg/bulnc/

[2] https://ifgpt.dcl.bas.bg/en/

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 12–24. Palma, Mallorca. May 11, 2026.

lows. First, we merge several large Bulgarian text collections into a unified dataset with standardised metadata and text formats. Second, we provide a unified metadata schema for all documents and organise the metadata categories in a graph-based representation. Finally, we offer efficient mechanisms for querying metadata to identify suitable documents for specific tasks such as LLM fine-tuning or Retrieval-Augmented Generation (RAG).

2. Bulgarian National Corpus and related work

The rapid growth of large-scale language data in recent years has advanced in several directions: expanding existing reference corpora with new text types, integrating multilingual and multimodal data, and producing training data specifically tailored for language technologies and large language models.

2.1. Corpus Query Tools

Most national corpora have an online search interface linked to a predefined document set. For some corpora, a dedicated corpus query tool has been developed, such as the PELCRA search engine for the National Corpus of Polish (Pęzik et al., 2016) and the KorAP corpus analysis platform (Diewald et al., 2016) for the German Reference Corpus DeReKo, among others.

Other corpora use open-source corpus query tools: KonText (Machálek, 2020) is a web-based corpus query tool for working with texts in the Czech National Corpus. The Slovak National Corpus is accessed via the NoSketch Engine (Rychlý, 2007; Kilgarriff et al., 2014). The Corpus of Written Standard Slovene is available via Sketch Engine, NoSketch Engine, and KonText (Krek et al., 2020).

In many corpus query tools, users can select which corpus to use and create their own collections of texts within the corpora by filtering according to various criteria, such as topic, subgenre, author, or source, and then search this subcorpus as if it were their own corpus. Tools such as Sketch Engine, NoSketch Engine, KonText, and AntConc (Anthony, 2024) allow access to multiple corpora, either as a single option or as a selection of subcorpora.

The Bulgarian National Corpus has a web interface for searching the corpus,[3] building concordances, and extracting examples (Koeva et al., 2012, 100-101). The search system allows complex linguistic queries involving different levels of annotation combined in various ways. It was designed to support both monolingual and parallel corpora in a uniform way. Compared to CQL, the implemented Designed Query Language (DQL) supports terms such as word, relation (i.e. word form, synonym, hypernym, etc.), and their combinations. Both ordered and unordered queries are supported, as well as conjunction and disjunction of ordered queries, such as searching for paraphrases.

[3] http://search.dcl.bas.bg

2.2. Expanding large reference corpora

Large reference corpora have increased in both volume and the range of text types they represent. Egbert et al. (2022) provide a systematic methodological overview of large corpora, arguing that size alone does not guarantee representativeness and proposing a two-pillar framework: domain coverage and distributional balance. The authors criticise the assumption that "bigger is always better", redirecting efforts from scale to design.

Hashimoto and Nelson (2024) examine 709 corpus descriptions published between 2010 and 2019. The authors analyse sampling decisions and the methodological principles employed by recent large-scale expansions, which rely on growth in corpus size accompanied by transparent, principled sampling techniques (Egbert et al., 2022; Hashimoto and Nelson, 2024).

Over the past 10 years, the main effort in developing the Bulgarian National Corpus through various national and international projects has focused on collecting, cleaning, and enriching large datasets, which will be discussed in more detail in the following section.

2.2.1. Large collections of unstructured data

The dominant paradigm for LLM pre-training data is large-scale web harvesting, with Common Crawl serving as the primary raw source for most major datasets. Conneau et al. (2020) introduced the CC-100 dataset – approximately two terabytes of filtered monolingual text in one hundred languages derived via the CCNet pipeline. Subsequent efforts have focused on aggressive quality filtering and deduplication.

Gao et al. (2020) established the multi-source paradigm with The Pile – 825 GiB of English text from 22 curated subsets spanning books, code, scientific papers, and online discussion – showing that domain diversity substantially improves downstream generalisation.

Soldaini et al. (2024) released Dolma (3 trillion tokens across six source types), providing both the data and the complete processing toolkit. Nguyen et al. (2024) released CulturaX (6.3 trillion tokens in 167 languages) by merging and cleaning mC4 and OSCAR. Singh et al. (2024) produced the Aya Collection, aggregating 513 million instances across 114 languages through an open participatory science model.

The Bulgarian National Corpus was expanded with several domain-specific corpora from the

OPUS collection (Tiedemann, 2012). The largest of these are the EMEA corpus of administrative medical texts and the OpenSubtitles corpus (film subtitles) (Lison and Tiedemann, 2016).

2.2.2. Special purpose datasets

Alongside general-purpose pre-training corpora, an increasing body of work focuses on developing datasets for specific task types or application contexts.

Instruction tuning has become a particularly active area for special-purpose dataset construction. Chung et al. (2022) introduced the FLAN v2 collection, comprising more than 1,800 reformatted tasks. Zhou et al. (2023) developed LIMA, a set of 1,000 carefully hand-selected prompt–response pairs that produce a competitive instruction-following model. Taori et al. (2023) operationalised the self-instruct paradigm by generating 52,000 instruction examples from GPT-3 and releasing both the data and training code. For multilingual instruction data, Muennighoff et al. (2023) present the ROOTS corpus (1.6 TB of text across 46 natural and 13 programming languages), which underpins the BLOOM model. Similarly task-focused, Kocoń et al. (2025) document the CLARIN-PL infrastructure, including the MultiEmo sentiment corpus extended across eleven languages.

Together, these efforts mark a shift from passive data collection to active dataset design oriented towards particular tasks, ensuring prompt diversity and quality control.

Within the structure of the BulNC, the Diachronic Corpus of Bulgarian has been compiled to support research on the lexical and grammatical features of Bulgarian over time.[4]

The corpus contains texts totalling 1.1 million words from 1851 to recent years, divided into six time intervals (1851–1880; 1881–1910; 1911–1930; 1931–1950; 1951–1990; 1991–2021), and covers three domains: fiction, news, and science. The texts are sourced from various places, including scanned copies of periodicals, international databases (e.g. Gutenberg), and modern electronic databases for texts from 1990 onwards. The choice of domains was based on observations of domain coverage across time periods. Administrative and other types of texts are rare in the earlier periods and are therefore not included in the Diachronic Corpus.

2.2.3. Multilingual Large Datasets

The development of large-scale multilingual corpora has involved sharing annotation schemes, frameworks, and data processing pipelines.

The TenTen corpus family (Jakubíček et al., 2013) covers more than fifty languages, each with over ten billion words, and has progressively added genre and topic classification in its latest releases.

The ParlaMint project (Erjavec et al., 2024) uses a shared Parla-CLARIN scheme, Universal Dependencies annotation, and named-entity labels. This results in a comparable corpus in which political discourse can be studied across languages and political systems.

Hundreds of parallel text corpora are already available with Bulgarian as one of the languages, most of which can be downloaded from repositories such as ELG[5] and CLARIN.[6] The bilingual corpora mainly contain Bulgarian and English or other European language pairs, for example, Bulgarian – Modern Greek, Bulgarian – German, Bulgarian – French, Bulgarian – Italian, and Bulgarian – Spanish.

Bulgarian is included in even more multilingual corpora, some of which are sentence-aligned, enabling straightforward cross-lingual research. Many large multilingual corpora are created automatically from web sources (e.g. Common Crawl, Wikipedia), while others are compiled from institutional, parliamentary, subtitle, or legislative data.

3. IfGPT composition

The IfGPT dataset is a collection of datasets, primarily in Bulgarian but also in English, based on the Bulgarian National Corpus. As the structure of BulNC has been described in detail elsewhere (Koeva et al., 2012), we focus here only on the main components that currently comprise the IfGPT dataset. These are summarised in Table 1.

The dataset **Bulgarian MARCELL** consists of legislative documents divided into fifteen types (Váradi et al., 2020). The documents span from 1946 to 2023 and were extracted from the Bulgarian State Gazette, the official gazette of the Bulgarian government, which publishes documents from official institutions such as the government, the Bulgarian National Assembly, the Constitutional Court, and others. The Bulgarian dataset contains 25,283 documents categorised into eleven types: Administrative Court; Agreements; Amendments (legal acts); Conventions; Decrees; Decrees of the Council of Ministers; Directives; Instructions; Laws (legal acts); Memoranda; Resolutions. The dataset comprises approximately 45 million tokens and 3,281,000 sentences (as of the end of March 2021). Bulgarian MARCELL is part of a comparable corpus of national legislative documents for seven languages (Bulgarian, Croatian, Hungarian, Polish,

[4] https://dcl.bas.bg/bulnc/en/dostap/izteglyane/

[5] https://live.european-language-grid.eu/

[6] https://www.clarin.eu/

Dataset & Language(s)	Domains	Size	Format & annot.	Source & Licence
Bulgarian MARCELL BG 1946–2023	Legal (11 types: admin. court, agreements, amendments, conventions, etc.)	25K texts; 3.28M sents; 45M tokens	CoNLL-U+; morph., dep., NER, EuroVoc/IATE annotation	Bulgarian State Gazette Public Domain
Bulgarian CURLICAT BG	7 domains: Culture, Education, EU, Finance, Politics, Economics, Science	6K texts; 22.8M tokens	CoNLL-U+; JSON; full ling. annotation	BulNC; science sources: books, PhD theses; web CC-BY CC-BY-SA CC-BY-NC
Aligned and Normalised Parallel Data BG–EN	16 domains: General News, BG Presidency, Economics, Culture, Military, Politics, etc.	1.1M sent. pairs; 19.0M words (BG); 19.2M words (EN)	Sentence-aligned pairs; partly manual selection & correction	Web media; institutional websites Public Domain Various
General News in Bulgarian BG	185 domains	2.1M texts; 33.4M sents; 601M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (11.8K domains, 2.1M pages) Various
General News in English EN	185 domains	5.9M texts; 166.7M sents; 3.3B words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (324.5K domains, 5.9M pages) Various
News about the the Bulgarian Presidency in Bulgarian BG	185 domains	36.8K texts; 698K sents; 16.6M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (613 domains, 36.8K pages) Various
News about the Bulgarian Presidency in English EN	185 domains	12.3K texts; 292K sents; 8.8M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (663 domains, 12.3K pages) Various
General News in English from Bulgaria EN	General news (Bulgarian electronic media)	19.1K texts; 876K sents; 18.6M words	Sentence splitting; tokenisation; language detection; normalisation & cleaning	Web crawling (140 BG domains, 19.1K pages); predefined BG domain list Various
Filtered General News in English from Bulgaria EN	185 domains	19.1K texts; 237K sents; 5.5M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (140 BG domains, 19.1K pages); predefined BG domain list Various
Filtered News about the Presidency in English from Bulgaria EN	185 domains	1.4K texts; 20.6K sents; 504.6K words	JSON; metadata; automatic categorisation; normalisation & cleaning	Focused web crawling (55 BG domains, 1.4K pages) Various
Collection of Bulgarian Texts BG	4 styles; 42 domains: Adventure, Archaeology, Chemistry, Computers, Court, Culture, Ecology, Economics, etc.	66 collection files; 28.9M sents	Sentence splitting (bgLPC); no metadata; arbitrary sentence order	Internet Various
Collection of English Texts EN	4 styles; 13 domains: Court, Culture, Ecology, Economics, Health, History, etc.	45 collection files; 8.1M sents	Sentence splitting (bgLPC); no metadata; arbitrary sentence order	Internet Various
IfGPT – News BG Until 1990	News (historical)	5.5M texts; 270.5M tokens	OCR; LLM-assisted article separation & metadata extraction	Printed periodicals Various
IfGPT – Periodicals BG Until 1990	Periodicals (historical)	25K texts; 30M tokens	OCR and pagination and metadata extraction	Printed periodicals Various
IfGPT – New periodicals BG After 1990	Contemporary periodicals	4.1M texts; 4.4B tokens	Text & metadata extraction; LLM-assisted post-processing	Contemporary press Various
IfGPT – Books BG	General (books)	22K texts; 630M tokens	OCR and pagination; title-page metadata extraction	Printed books Various
Multilingual Image Corpus (MIC21) Image BG Various	4 domains (Sport, Transport, Arts, Security); 130 subdomains	22K images; 230M objects	annotated objects in images; short narrative descriptions	Open repositories CC-BY-SA

Table 1: Overview of the components in the IfGPT dataset with their key features. The Bulgarian language processing pipeline, is available at: http://dcl.bas.bg/dclservices/

Romanian, Slovak, and Slovenian) collected within the project **Multilingual Resources for CEF.AT in the Legal Domain** (**MARCELL**).[7] The Bulgarian MARCELL dataset is annotated in CoNLL-U Plus format (Koeva et al., 2020) for morphosyntax, dependency structure, and named entities. Documents are classified into thematic domains and enriched with specialised terminology identified from IATE[8] and EuroVoc.[9]

The dataset **Bulgarian CURLICAT** contains 113,087 documents divided into seven thematic domains: Culture, Education, European Union, Finance, Politics, Economics, and Science (Váradi et al., 2022a). The dataset comprises 6,036 documents with a total of 22,809,225 tokens. All documents are licensed under CC-BY, CC-BY-SA, or CC-BY-NC. To ensure a sufficient number of copyright-free documents, several sources were identified, including a library of scientific texts (books and PhD theses) and other websites providing texts from the required thematic domains. The texts are categorised, linguistically annotated, and provided in CoNLL-U Plus format, in the same way as the Bulgarian MARCELL dataset. The dataset was created within the CURLICAT project (**Curated Multilingual Language Resources for CEF.AT**) (Váradi et al., 2022b,a), which extended the approach to provide comparable corpora in domain-specific areas for the same seven languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, and Slovenian.[10]

Within the project **CEF Automated Translation for the EU Council Presidency**,[11] a dataset of **Aligned and Normalised Parallel Data** has been collected, curated and annotated, comprising the thematic domains presented in Table 2. These include parallel texts in English and Bulgarian aligned at sentence level. The parallel texts have been selected and collected from electronic media: Bulgarian National Radio, Bulgarian National Television, Bulgarian News Agency, Focus Information Agency, Sofia News Agency, web publications, web newspapers, and institutional websites with open access to relevant texts in Bulgarian and English.

The dataset **General News in Bulgarian** contains news from various thematic domains. The news and metadata were automatically acquired from different, predominantly Bulgarian, internet sources: 11,840 web domains and 2,116,739 web pages. The total number of words in the collected General News in Bulgarian is 601,330,975, distributed across 33,375,366 sentences. A crawling platform was used to identify and acquire monolingual data from websites, detect near-duplication at the document level, and normalise and clean the text. The resulting texts were structured into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 thematic domains, ordered by probability. The distribution of texts across the most frequent thematic domains is shown in Table3.

Domain	Sentence pairs	Words (BG)	Words (EN)
General News	3,118	61.9K	64.6K
BG Presidency	3,000	69.3K	74.6K
Science	454	9.2K	9.5K
Fiction	1,177	12.6K	13.3K
Economics	10.1K	201.6K	199.9K
Culture	3,164	54.6K	56.4K
Military	12.7K	265.8K	255.5K
Politics	66.2K	1.3M	1.3M
Social Domain	578	11.6K	11.6K
Undetermined	224.9K	4.6M	4.5M
Ecology	238.7K	4.9M	4.8M
History	5,215	98.6K	108.1K
Philosophy	725	15.3K	18.2K
Psychology	852	20.6K	25.2K
Medicine	511.4K	6.7M	7.0M
Technology	50.2K	712.6K	703.2K
Total	**1.1M**	**19.0M**	**19.2M**

Table 2: Overview of parallel texts in the **Aligned and Normalised Parallel Data** by domain

Domain	Number of texts
Economy	919,599
Sociology	735,566
Politics	718,540
Law	711,131
Enterprise	591,679
Commerce	455,647
Pedagogy	447,613
Administration	410,567
School	397,465
Free Time	365,550
History	356,974

Table 3: Distribution of texts in the most frequent domains in the **General News in Bulgarian** dataset.

The dataset **General News in English** contains news from various thematic domains. The news articles, together with some metadata, were automatically collected from numerous internet sources: 324,493 web domains and 5,961,124 web pages. The total number of words in the collected General News in English is 3,324,746,119, distributed across 166,718,125 sentences. A crawling platform was used to identify and acquire monolingual data from websites, detect near-duplication at the document level, and normalise and clean the text. The resulting texts were structured into JSON files,

[7] https://marcell-project.eu
[8] https://iate.europa.eu
[9] https://eur-lex.europa.eu
[10] https://curlicat.eu
[11] https://tilde.ai/machine-translation/

which contain extracted metadata and automatic categorisation of the content into 185 thematic domains (ordered by probability). The distribution of texts across the most frequent thematic domains is shown in 4.

Domain	Number of texts
Economy	4,516,499
Enterprise	3,895,116
Commerce	3,797,299
Exchange	2,879,980
Law	2,764,209
Finance	2,592,014
Bookkeeping	1,720,943
Banking	1,632,725
Politics	1,608,819
Sociology	1,597,035

Table 4: Distribution of texts in the most frequent domains in the **General News in English** dataset.

The dataset **News in Bulgarian about the Bulgarian Presidency of the Council of Europe** contains news thematically related to the Bulgarian Presidency. The news articles, together with some metadata, were automatically collected from various sources: 613 web domains and 36,835 web pages. The total number of words in the collected Bulgarian news is 16,550,562, distributed across 698,434 sentences. A crawling platform was used to acquire monolingual data from websites, as well as for normalisation, cleaning, and near-duplicate removal at the document level. The texts were aggregated into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 domains. The thematic domains most frequently assigned to the largest number of JSON files are: politics (27,089 files), economy (25,688 files), sociology (24,941 files), law (20,656 files), enterprise (19,526 files), administration (19,358 files), history (15,052 files), and diplomacy (11,167 files).

The dataset **News in English about the Bulgarian Presidency of the Council of Europe** contains news thematically related to the Bulgarian Presidency. The news articles, along with some metadata, were automatically acquired from various sources: 663 domains and 12,327 pages. The total number of words in the collected news in Bulgarian is 8,794,285, distributed across 292,111 sentences. A crawling platform was used to acquire monolingual data from websites, as well as for normalisation, cleaning, and near-duplication detection at the document level. The texts were aggregated into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 domains. The domains most frequently assigned to the largest number of JSON files are: economy (5,590 files), politics (5,531 files), enterprise (4,600 files), law (4,488 files), sociology (4,484 files), administration (4,011 files), diplomacy (3,700 files), finance (3,464 files), history (2,953 files), commerce (2,672 files), and exchange (2,287 files).

The following three datasets are sourced specifically from the Bulgarian web domain. Filtering between Bulgarian and non-Bulgarian web domains was conducted by extracting the country code from the WHOIS (BG) database, identifying the IP GeoLocation, and manually filtering the list of domains containing specific words, such as the names of Bulgarian cities.

The dataset **General News in English from Bulgaria** was automatically collected from various sources in Bulgaria: 140 web domains and 19,120 web pages. The total number of words in the collected General News in English is 18,631,384, distributed across 876,739 sentences.

The dataset **Filtered General News in English from Bulgaria** is drawn from the same 140 web domains and 19,120 web pages. However, the total number of words is lower: 5,512,392, distributed across 237,371 sentences, as the texts are further filtered to ensure that their content is specifically focused on Bulgaria.

The dataset **Filtered News in English about the Bulgarian EU Council Presidency from Bulgaria** was collected from 55 web domains and 1,402 web pages. The total number of words in the collected news in English is 504,596, distributed across 20,616 sentences.

The texts in the three datasets were aggregated into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 domains. Automatic linguistic processing – sentence splitting and tokenisation – was performed.

The dataset **Collection of Bulgarian Texts** contains texts obtained mainly from the internet. Texts are organised into 66 files by style and thematic domain. A general classification into different styles (administrative, science, news, and fiction) is provided, and texts are further classified into thematic domains: Adventure, Archaeology, Architecture, Arts, Astronomy, Biology, Chemistry, Children, Computers, Court, Culture, Ecology, Economics, Education, Engineering, Entertainment, Geography, Health, History, Law, Lifestyle, Linguistics, Literature, Maths, Medicine, Military, Pedagogy, Philosophy, Physics, Politics, Psychology, Relations, Religion, Science Fiction, Science, Society, Sociology, Sport, Technology, Travel, and Unclassified. Automatic sentence splitting is applied; the text format is one sentence per line. The collection contains 28,919,379 sentences. The files are not supplemented with metadata and the sentence order is arbitrary.

The dataset **Collection of English Texts** contains texts obtained mainly from the internet. Texts are organised into 45 files by style and thematic do-

main. A general classification into different styles (administrative, science, news, and fiction) is provided, and texts are further classified into thematic domains: Court, Culture, Ecology, Economics, Health, History, Military, Physics, Politics, Science Fiction, Society, Technology, and Unclassified. Automatic sentence splitting is applied; the text format is one sentence per line. The collection contains 8,144,881 sentences. The files are not supplemented with metadata and the sentence order is arbitrary.

The project **Infrastructure for Fine-tuning Pretrained Large Language Models** (IfGPT)[12] provides an opportunity to collect, clean, and curate additional Bulgarian data. Table 5 shows the amount of new data collected. This includes older texts, both periodicals and books, which require further processing – OCR, pagination, and metadata extraction using the layout and structure of the texts (for example, metadata such as date and source in the header and footer of periodicals, publishing information from the title page, etc.)—as well as post-processing procedures for the extracted text, including the removal of boilerplate content and correction of hyphenation. The capabilities of LLMs have been tested for one or more of these tasks; in particular, OCR was combined with text completion (separating articles in newspapers) and metadata extraction using the Claude Sonnet 4.6 API.

Source	# texts	# tokens	Licence
News up to 1990	5,544K	270,52M	various
Periodicals up to 1990	25K	30M	various
New periodicals	4,119K	4,378M	various
Books	22K	630M	various

Table 5: New data added to IfGPT dataset.

The **Multilingual Image Corpus** (**MIC21**) is a recently developed dataset designed to advance research in multilingual and multimodal data processing (Koeva et al., 2022). It provides pixel-level annotations for over 203,000 objects in more than 21,000 images, covering 730 classes organised into four thematic domains and 130 subdomains.[13] These annotations support the development of specialised models for object detection, segmentation, and classification. The annotated object classes depicted in the images are structured within an Ontology of Visual Objects, enabling the construction of diverse datasets for a wide range of tasks. This ontological framework supports learning inter-object associations, identifying relationships between objects, and aligning objects and their relations with textual content. Class labels are enriched with synonyms, definitions, and usage examples in 25 languages, making the dataset suitable for applications such as multilingual image captioning, visual question answering, and multimodal machine translation. In a recent extension of the dataset, images have been accompanied by brief narrative de

[12] https://ifgpt.dcl.bas.bg/en/

[13] https://dcl.bas.bg/en/projects_list/mic21/

4. Extensive metadata description in IfGPT

Each document in the IfGPT dataset is described by a set of mandatory and optional metadata fields. The mandatory fields ensure consistent descriptions across all documents, covering text characteristics, domain information, and quantitative document statistics. The optional fields provide supplementary descriptive details where available, including authorship, stylistic properties, and task suitability. Most of the metadata originates from the BulNC metadata, which was structured as a graph (Koeva et al., 2016); however, there are some specific metadata fields related to LLMs, for example: **PersonallyIdentifiableInformation**, **BiasedInformation**, **LicenseLink**,**TaskCategories**.

4.1. Metadata types

The mandatory metadata includes: **Identifier** (unique document ID with language prefix `bg`), **Licence** (terms of use; various CC licence types, etc.), **PublicationDate** (original publication date), **DocumentTitle** (title of the document), **Source** (journal or website), **Medium** (modality: text, multimodal), **Url** (original web address), **Domain** (up to six thematic domains from a predefined list), **Keywords** (up to six descriptive terms), **NumberWords** (total word count), **NumberSentences** (total sentence count), **NumberParagraphs** (total paragraph count), **NumberTokens** (total token count), **PersonallyIdentifiableInformation** (an array of values for all sentences, calculated as the proportion of tokens flagged as PII), and **BiasedInformation** (an array of values for all sentences, calculated as the proportion of tokens flagged as potentially expressing bias).

The optional metadata includes: **Author** (name(s) of the text's creator(s)), **Style** (literary register, e.g. Legal, Journalism, Administrative), **Type** (document genre, e.g. book chapter, newspaper article, blog post), **Subdomain** (narrower thematic classification linked to a parent domain), **TranslatedDocument** (whether the document is original Bulgarian or a translation), **CollectionDate** (date of data collection in ISO 8601), **LicenseLink** (URL of the licence describing its terms), and **TaskCategories** (intended NLP applications from a predefined list, e.g. question answering).

An illustrative example of metadata organisation is shown on Figure 1.

4.2. Metadata management

To store and manage the metadata described above, a Neo4J graph database is used.[14] Graph databases are well suited to this purpose, as they are designed to handle large volumes of interconnected data efficiently, support horizontal scaling, and maintain performance even under complex queries (Francis et al., 2018). Neo4J is chosen for its high performance, native support for the Cypher query language, and extensive community ecosystem.

The metadata is organised according to a graph schema that captures the key entities and their interrelationships. Five node types are defined: **Document** nodes, which contain the core descriptive properties of each text (identifier, title, source, domain, author, licence, etc.); **Domain** nodes, characterised by a name and a parent category to support hierarchical thematic classification; **Author** nodes, storing author names and optional biographical details; **Source** nodes, recording the name and URL of the publishing organisation; and **Licence** nodes, defined by a single type property.

The relationships between these nodes are represented as directed edges: a **Document** belongs to a **Domain** (`BELONGS_TO`); a **Domain** may be a subcategory of another **Domain** (`SUBCATEGORY_OF`); a **Document** is licensed under a **Licence** (`LICENSED_WITH`); a **Document** is attributed to an **Author** (`WRITTEN_BY`); a **Document** is associated with a **Source** (`PUBLISHED_IN`); etc. Together, these nodes and edges form a flexible, queryable representation of the metadata that supports both document retrieval and downstream NLP tasks.

The metadata graph database currently contains a total of 237,795 documents, described through `metadata`, `Author`, `Domain`, and `Licence` nodes, connected by a number of relation types: `BELONGS_TO`, `LICENSED_WITH`, `WRITTEN_BY`, `PUBLISHED_IN`, etc.

The use of a graph database for managing metadata offers several key advantages. It models the rich relations interconnecting documents, domains, licences, sources, and more, making it much more expressive for metadata exploration. Complex queries combining multiple criteria are handled as efficient graph traversals (Figure 2 illustrates this process). Graph databases also scale efficiently when new relationship types are introduced (e.g. linking Keywords and Domain nodes) without requiring changes to the schema.

Domain distribution. The collection spans 45 thematic domains and subdomains. The most represented domains account for the majority of documents: Science (19.4%), Public Administration (15.9%), Economics (15.4%), and Politics (12.7%); 16.7% of texts have no assigned domain. The remaining 41 domains each account for less than 6% of documents, with several highly specialised domains, such as Architecture, Computer Science, and Physics.

Licensing. The vast majority of documents are openly licensed (93.4%), with the most common licences being CC-BY-SA (42.8%) and CC0 (45.9%). Only 6.6% of documents have a restricted licence.

Authorship. Author metadata is available for 93.7% of documents, while 6.3% lack authorship information. The collection references 4,563 distinct authors across all documents, including both individuals and organisations.

Document statistics. The collection currently available for searching comprises approximately 718.4 million words and 866 million tokens. On average, each document contains 3,642 tokens (3,021 words) in 207 sentences, indicating that the collection predominantly consists of full-length texts.

Time period coverage. Publication dates range from 1935 to 2022. The bulk of the collection is concentrated in the period 2008–2011, which accounts for approximately 55% of all documents. Current efforts aim to provide more data covering the pre-2000 (pre-digital) period.

5. Ensuring linguistic integrity of data

Ensuring the linguistic integrity of the data is a major priority in compiling the IfGPT dataset, as poor content quality has been shown to directly affect model performance and introduce systematic biases (Kreutzer et al., 2022). To address this, we have implemented two main procedures: deduplication and the removal of unsuitable and malformed texts.

Deduplication is carried out in two steps. First, metadata attributes such as source, publication year, domain, title, and author are used to identify and remove identical texts appearing in multiple text collections, which substantially reduces the computational burden of subsequent processing. The core deduplication procedure is then performed using MinHash combined with Locality Sensitive Hashing (LSH) (Lee et al., 2022), which detects both exact and near-duplicate texts by estimating N-gram overlap across document pairs.

The second procedure involves identifying and removing boilerplate (e.g., navigation menus, repeated footer text, legal disclaimers), correcting typographical errors, removing incomplete or malformed sentences, and filtering harmful, offensive, and toxic content.

[14] https://neo4j.com/

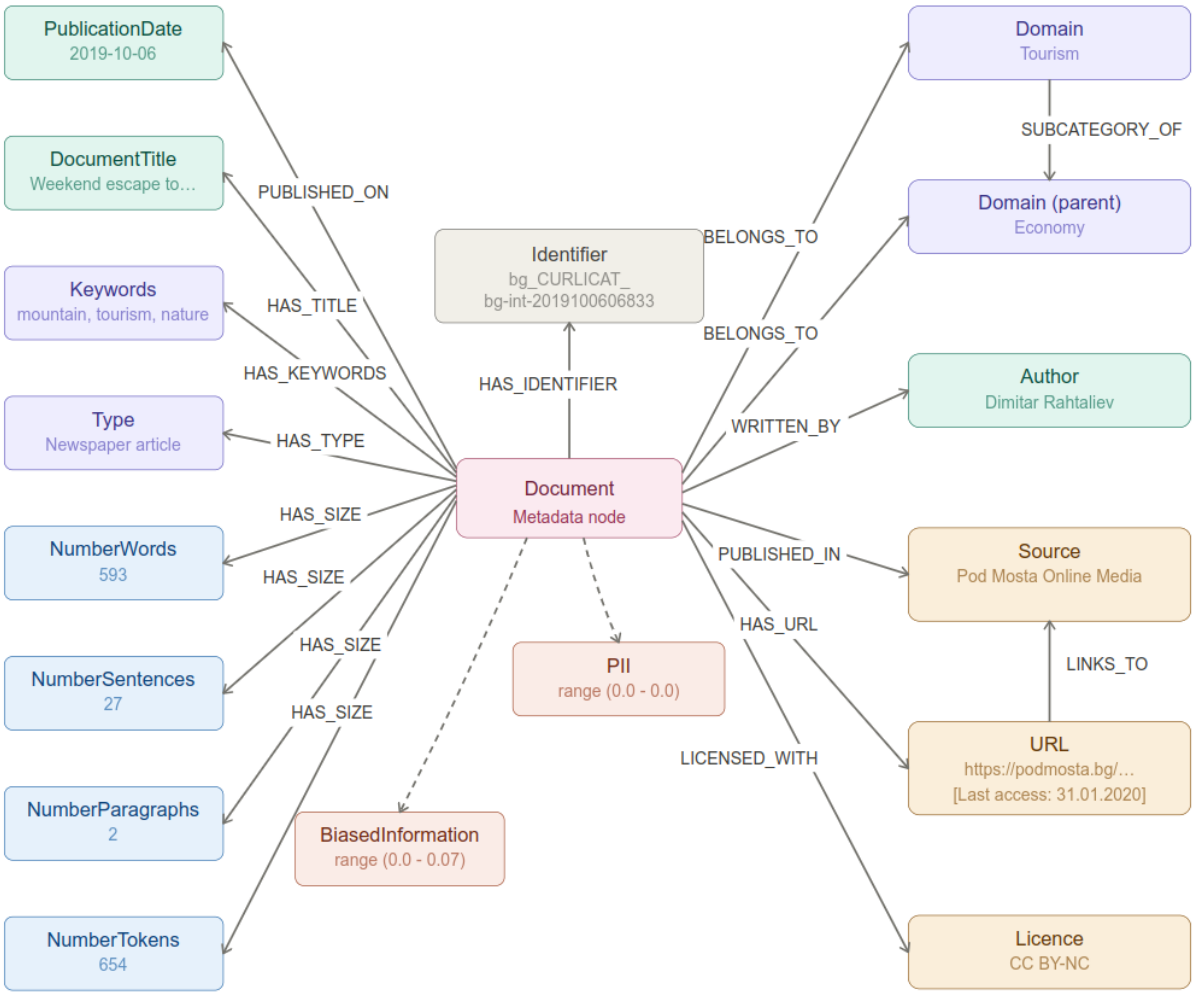


Figure 1: IfGPT Dataset Metadata.

The PII and BiasedInformation metadata fields are also directly relevant to the quality assessment and maintenance of the data. Instead of being discarded, potentially sensitive information and biased content are flagged, allowing users to apply their own filtering criteria depending on the task at hand.

6. Access

The IfGPT dataset is publicly accessible through a dedicated web interface,[15] which allows users to search and browse the complete collection of documents. The interface offers four filtering mechanisms (see Figure 3).

Licence filter. Users can restrict results by licence type, selecting either general or specific licences, e.g. all Creative Commons or specific licences, other open licences, and can include or exclude data with restricted licences.

Domain filter. Documents can be filtered by one or more domains.

Period filter. Users can specify a publication date range by setting a start and/or end year to restrict results to a particular period (omitting either sets it to the default, i.e. the earliest or latest document year in the database).

Keywords filter. Free-text keyword search is supported, with multiple keywords accepted as a comma-separated list.

Search results are displayed as paginated document cards, each showing the document title, domain tags, licence, document type, publication date, source, URL to the original source, and quantitative properties including the number of paragraphs, sentences, and words (see Figure 3).

The interface also provides three download options for the retrieved results: (1) metadata of retrieved documents (JSON format), (2) list of links to original sources (TSV format), and (3) full data (link to a ZIP archive) subject to confirmation of details and agreement to the Terms and Conditions for download, including the restrictions imposed by the licences of the original documents.

[15] Available at https://ifgpt.dcl.bas.bg/ifgpt-dataset/

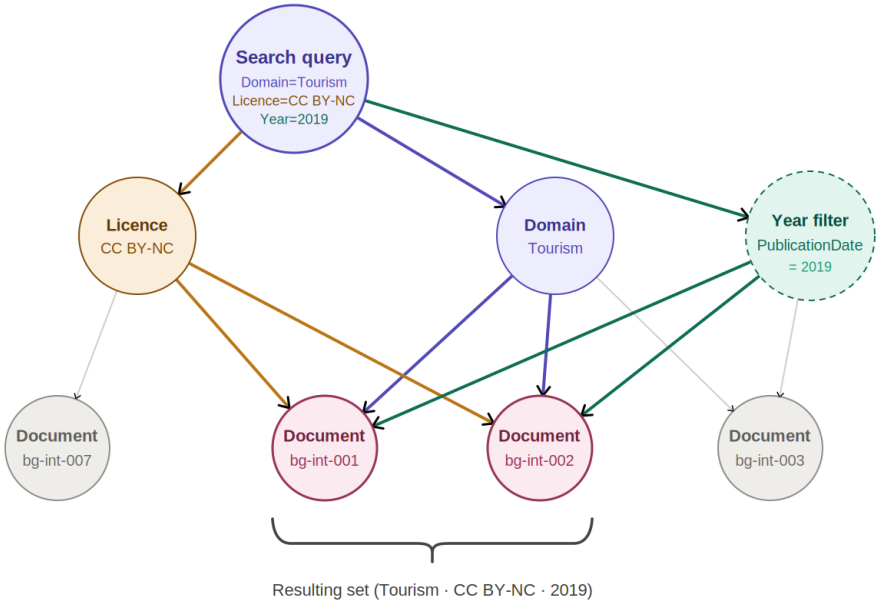


Figure 2: IfGPT Dataset Search in the graph database.

7. Conclusion

By describing the IfGPT dataset, we demonstrate that the reference corpora are suitable for inclusion in LLM datasets, as they share several important similarities in design, compilation, and management. Both BulNC and IfGPT aim to represent authentic Bulgarian language and include multiple genres, domains, and text types, as both linguistic research and language model training require exposure to varied language. Both resources also require cleaning and deduplication, filtering out incomplete or malformed texts. Both are stored in standard formats, divided into documents, paragraphs, and sentences. Although the IfGPT dataset does not provide as extensive linguistic annotation as BulNC, such annotation could also be implemented. This means that both resources can complement each other: the BulNC is part of the IfGPT dataset, but appropriate parts of the IfGPT dataset that fulfil balance and distribution requirements can also be added to the Bulgarian National Corpus. Intensive work on developing tools for detecting bias and PII in IfGPT texts will also be useful for application to BulNC documents.

Unlike some LLM datasets, the IfGPT dataset, inheriting from BulNC, possesses extensive metadata descriptions. The BulNC metadata, originally organised as a graph, is enhanced with some LLM-specific metadata such as PII and bias scores, licence information, and is further managed as a graph database. The common metadata scheme for both resources is beneficial in two ways: searching through the metadata for relevant texts from IfGPT for BulNC, and adding new relevant texts simultaneously to BulNC and IfGPT.

Future developments include: (a) adding new and diverse text data to both resources; (b) expanding metadata descriptions, especially for texts for which some metadata categories are not assigned values; (c) validating and improving text data quality in both resources; and (d) enhancing accessibility and providing easy access to the IfGPT data for various purposes.

8. Acknowledgments

The present study is carried out within the project Infrastructure for Fine-tuning Pretrained Large Language Models, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Laurence Anthony. 2024. AntConc (Version 4.3.1) [Windows, macOS, Linux]. Corpus analysis toolkit.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *ArXiv*.

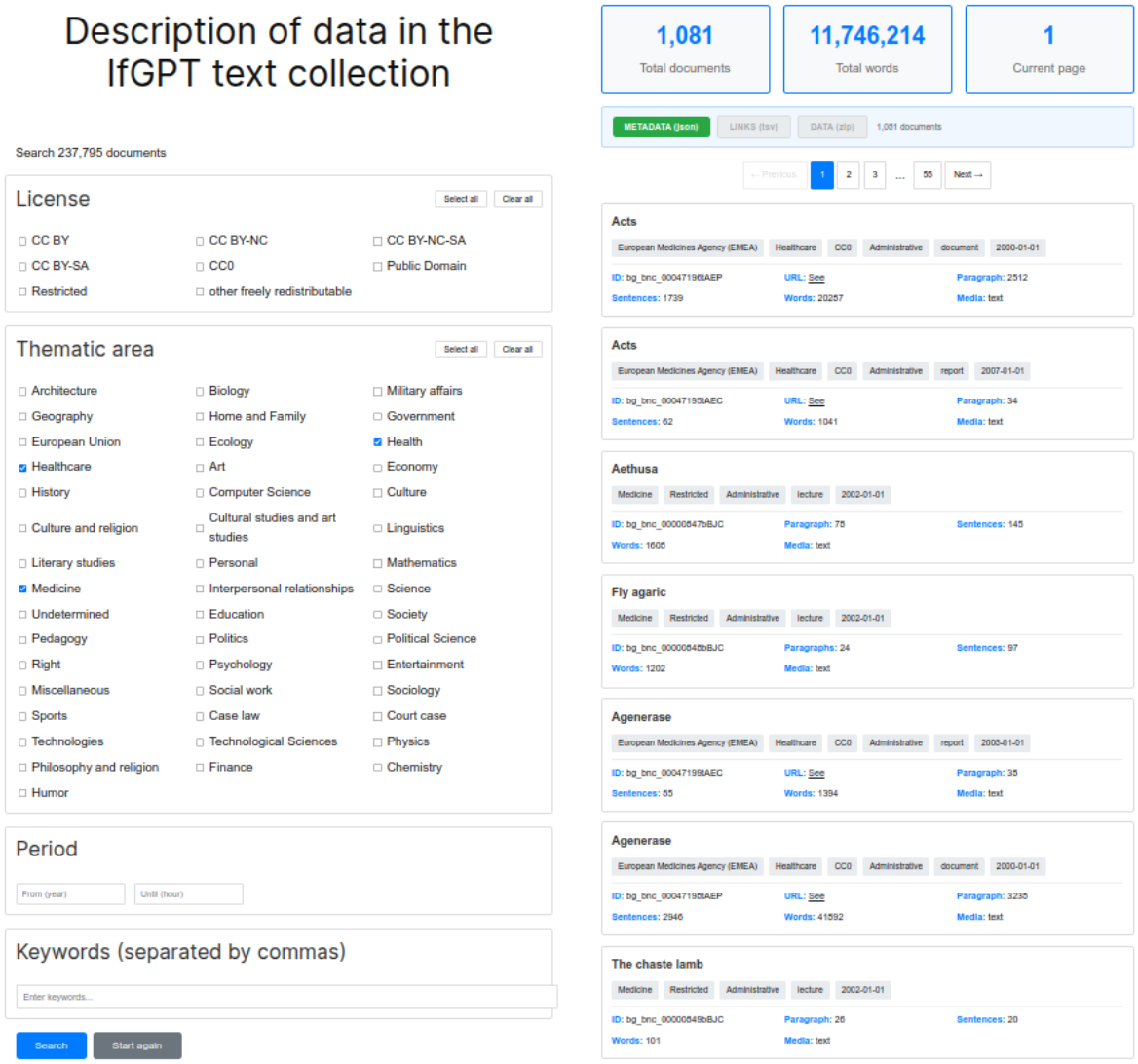


Figure 3: Web interface for searching and selecting datasets from IfGPT. Results from the search are shown on the right

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP architecture – diving in the deep sea of corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3586–3591, Portorož, Slovenia. ELRA.

Jesse Egbert, Douglas Biber, and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge University Press, Cambridge.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darundefinedis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2024. ParlaMint II: advancing comparable parliamentary corpora

across Europe. *Language Resources and Evaluation*, 59(3):2071–2102.

Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An Evolving Query Language for Property Graphs. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1433–1445, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*.

Brett J. Hashimoto and Mike Nelson. 2024. Recent trends in corpus design and reporting: A methodological synthesis. *Research in Corpus Linguistics*, 12(1):59–88.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen corpus family. Lexical Computing Ltd. / Masaryk University, Brno.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1(1):7–36.

Jan Kocoń, Mateusz Kopeć, et al. 2025. CLARIN-PL: A user-centred language technology infrastructure. *Language Resources and Evaluation*.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. ELRA.

Svetla Koeva, Ivelina Stoyanova, and Jordan Kralev. 2022. Multilingual image corpus – towards a multimodal and multilingual dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1509–1518, Marseille, France. ELRA.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 1(1):65–110.

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva, and Tsvetana Dimitrova. 2016. Metadata extraction, representation and management within the Bulgarian National Corpus. In *Priceedings of the 4th Workshop on Challenges in the Management of Large Corpora*, pages 33–39. ELDA.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: The reference corpus of written standard Slovene. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3340–3345, Marseille, France. ELRA.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, Portorož, Slovenia. ELRA.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings*

of the 12th Language Resources and Evaluation Conference (LREC 2020), pages 7003–7008, Marseille, France. ELRA.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and benchmarking of LLM agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6129–6139, New York, NY, USA. Association for Computing Machinery.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Leni Fowl, et al. 2023. ROOTS: A multilingual annotated pretraining corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 369–380, Toronto, Canada. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and pragmatic multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.

Piotr Pęzik, Paweł Kowalczyk, Łukasz Dróżdź, and Paweł Wilk. 2016. PELCRA for National Corpus of Polish: Search Engine 2. CLARIN-PL Digital Repository.

Pavel Rychlý. 2007. Manatee/Bonito: A Modular Corpus Manager. In *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, pages 65–70.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Ian Magnusson, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. Stanford.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey. ELRA.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiş, Dan Tufiş, Radovan Garabík, Simon Krek, Andraž Repar, Matjaž Rihtar, and Janez Brank. 2020. The MARCELL legislative corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France. ELRA.

Tamás Váradi, Bence Nyéki, Svetla Koeva, Marko Tadić, Vanja Štefanec, Maciej Ogrodniczuk, Bartłomiej Nitoń, Piotr Pęzik, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Dan Tufiş, Radovan Garabík, Simon Krek, and Andraž Repar. 2022a. Introducing the CURLICAT corpora: Seven-language domain specific annotated corpora from curated sources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 100–108, Marseille, France. ELRA.

Tamás Váradi, Marko Tadić, Svetla Koeva, Maciej Ogrodniczuk, Dan Tufiş, Radovan Garabík, Simon Krek, and Andraž Repar. 2022b. Curated multilingual language resources for CEF AT (CURLICAT): Overall view. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 341–342, Ghent, Belgium. European Association for Machine Translation.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Merimënga: A Manifest-First Pipeline for Reproducible Albanian Web Corpus Construction

Besim Kabashi and Michael Ruppert
Computational Linguistics, University of Tuebingen
besim.kabashi@uni-tuebingen.de

Computational Corpus Linguistics, FAU Erlangen-Nuremberg
michael.ruppert@fau.de

Abstract

We present Merimënga, a pipeline for reproducible Albanian web-corpus construction from Common Crawl. Rather than distributing a static text dump, we publish versioned manifests and append-only JSONL ledgers that make every retrieval and filtering decision replayable at record level. Records are addressed by (WARC filename, byte offset, byte length) and retrieved via HTTP range requests with checksum validation, enabling selective download, resumability, and exact re-materialization. On top of deterministic cleaning and deduplication, Merimënga supports teacher–student filtering: a large LLM labels a stratified sample; the resulting policy is distilled into a faster student model applied at corpus scale. The paper contributes (i) a reproducibility specification for web-corpus construction based on coordinate-addressed retrieval and decision ledgers, (ii) a concrete instantiation for Albanian with language-specific filtering, and (iii) an evaluation protocol for rerun equivalence and filter-stack ablation. Large-scale download and full-corpus filtering are ongoing; this submission focuses on methodology and auditable artifacts rather than final corpus statistics.

Keywords: Common Crawl, Reproducibility, Corpus Construction, Learned Filtering, Albanian

1. Introduction

Large web corpora are easy to grow, but much harder to reproduce, audit, and share responsibly. For lower-resource languages such as Albanian, these issues are amplified by domain concentration, noisy page structure, and practical licensing constraints around redistribution. Yet the need for such corpora is pressing: Albanian NLP, lexicography, and corpus-linguistic research all depend on large, clean text collections that are currently available only through a small number of multilingual releases with limited control over construction decisions.

Our approach combines two ideas: (1) recording every retrieval and filtering decision in append-only ledgers so that the entire construction process can be replayed from machine-readable artifacts, and (2) using LLM-based teacher–student filtering to scale quality decisions. Compared with CCNet/RefinedWeb-style quality pipelines (Wenzek et al., 2020; Penedo et al., 2023), the methodological contribution is the reproducibility framework itself, and the primary output is a replayable construction process rather than a static text release.

This design addresses a gap that is especially relevant for lower-resource languages. Large multilingual projects like OSCAR and HPLT provide invaluable data, but their release cycles are tied to project timelines, and replicating their pipelines independently requires comparable infrastructure. Merimënga targets a different use case: a research group working on a single language can build and extend a corpus from Common Crawl on modest hardware, inspect every filtering decision, adjust parameters, and incorporate new crawl snapshots incrementally—without depending on the release schedule of a multilingual project.

2. Problem Setting and Goals

The project addresses three concrete requirements that are common in large-corpus management:

1. **Reconstruction requirement:** a third party should be able to replay corpus construction decisions from machine-readable artifacts.

2. **Operational requirement:** long runs must survive partial failures, rate limits, and interrupted execution without losing state.

3. **Governance requirement:** publication should support reuse while reducing redistribution risk where possible.

These requirements motivate the core design choices described in the following sections.

3. Related Work and Positioning

OSCAR and related Common-Crawl pipelines established large multilingual releases and evolved substantially across versions (Ortiz Suarez et al., 2019; Abadji et al., 2021, 2022; OSCAR Project,

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 25–31. Palma, Mallorca. May 11, 2026.

Dimension	OSCAR-style releases	Merimënga
Primary objective	release-oriented corpus product	reproducible construction process
Primary artifact	release corpus + metadata	manifests + code + ledgers
Audit granularity	release version level	per-record decision trail
Retrieval model	download full release	fetch individual records by coordinate
Redistribution strategy	text release policies	manifest-only by default

Table 1: Positioning of our approach relative to OSCAR-style corpus releases.

2019, 2022, 2023; Brack et al., 2024). Later OSCAR versions move toward document-oriented releases with richer metadata and quality signals (OSCAR Project, 2021, 2022, 2023). More recent web-corpus projects (CCNet, mC4, RefinedWeb, FineWeb, DCLM, HPLT) emphasize strong filtering, deduplication, and ablation-driven development (Wenzek et al., 2020; Xue et al., 2021; Penedo et al., 2023, 2024; Li et al., 2024; Samuel et al., 2024). Teacher-student filtering specifically is present in FineWeb-Edu, which uses LLM-generated supervision to train a scalable quality classifier (Penedo et al., 2024; Hugging Face, 2024).

These projects share a common pattern: they process the full Common Crawl archive centrally, across all languages, and publish the result as a versioned release. While several of them release their code, replicating the full pipeline independently requires comparable infrastructure and access to the same crawl data. In practice, this means that (a) release schedules are tied to project capacity, and (b) the barrier to running such a pipeline for a single language is disproportionately high. Merimënga addresses a different point in this design space: it enables targeted, single-language corpus construction that is fully replayable, incrementally extensible, and feasible on modest hardware. The filtering techniques we employ (language identification, deduplication, learned quality scoring) are not novel in themselves; our contribution is the framework that makes their application reproducible and auditable at record level.

4. Corpus Generation and Distribution

This section describes how records are selected from Common Crawl and how the resulting corpus is distributed.

4.1. Selective Retrieval from Common Crawl

Rather than processing an entire Common Crawl snapshot, the pipeline queries Common Crawl's index for Albanian records, retrieves only those, and applies staged cleaning and filtering. Common Crawl is published in standardized archive formats (WARC and derived representations) (Common Crawl Foundation, 2026a; International Organization for Standardization, 2017).

Selection is index-based via Common Crawl index services. Index rows provide file path, byte offset, and record length, enabling targeted retrieval through HTTP range requests instead of full-segment downloads (Common Crawl Foundation, 2026b). This is a form of selective sampling: rather than downloading and processing an entire crawl snapshot (tens of terabytes), we query the index for Albanian pages only (`language=sqi, status=200, mime=text/html`) and retrieve just the matching records by their byte coordinates. Albanian accounts for roughly 0.05% of the Common Crawl archive (for comparison, OSCAR 23.01 lists approximately 497 k Albanian documents out of 1.13 billion total; HPLT v1.2 provides 1.24 million cleaned Albanian documents totalling 1.34 billion words (Samuel et al., 2024)). This reduces the data volume by several orders of magnitude compared to full-crawl processing, making single-language corpus construction feasible on a single workstation without cloud infrastructure. The same approach scales to larger setups—on a more capable server, multiple languages can be processed in parallel—but the entry barrier for a single low-resource language is low. This stands in contrast to projects like OSCAR or HPLT, which process the full archive across all languages simultaneously and therefore require substantial compute resources (Ortiz Suarez et al., 2019; Samuel et al., 2024).

For archive parsing and response extraction, we follow standards-compatible handling of WARC/HTTP records (Webrecorder, 2026; International Organization for Standardization, 2017). Deterministic reconstruction is supported by pinning each crawl snapshot (for example `CC-MAIN-2026-08`) and persisting retrieval coordinates with provenance metadata.

A limitation of this index-based sampling is that it depends on the language labels assigned by Common Crawl's own classifier. Pages that are Albanian but mislabeled (or unlabeled) in the index will be missed. We accept this trade-off in exchange for the practical efficiency of targeted retrieval, and note that the same limitation applies to any index-based approach.

Once records are retrieved, they pass through a multi-stage filtering stack—heuristic cleaning, language identification, deduplication, perplexity scor-

ing, and learned quality filtering—described in detail in Sections 5–6. Because Common Crawl is heterogeneous and noisy, we prioritize quality over volume, following established curation strategies (Wenzek et al., 2020; Ortiz Suarez et al., 2019; Abadji et al., 2022).

We release code, manifests, and decision ledgers rather than extracted full text. Downstream users retrieve content directly from Common Crawl and remain responsible for compliance with applicable copyright and terms (Common Crawl Foundation, 2024; Schäfer and Bildhauer, 2016). Reproducibility holds as long as the referenced snapshots remain publicly available (Common Crawl Foundation, 2026c,b). The pipeline design is language-agnostic: adapting it to another language requires replacing the language code in the index query and the language-specific filter resources. Because retrieval and filtering across multiple snapshots are still ongoing, we do not report a final corpus size in this version.

5. Pipeline Overview

The pipeline is implemented in Python 3.10. The retrieval and manifest-management components use only the standard library; filtering stages additionally rely on FastText for language identification (Joulin et al., 2017), KenLM for n-gram perplexity scoring (Heafield, 2011), and llama.cpp with LM Studio for local LLM inference in the teacher-labeling stage.

Crawl snapshots are processed in reverse chronological order, starting with the most recent and working backwards. This prioritizes fresh content and allows incremental expansion: each additional snapshot contributes only records not yet seen, and cross-snapshot deduplication keeps the most recent version of each page by default. Manifests have been generated for 19 crawl snapshots covering a total of 1.87 million candidate Albanian records. Download of the most recent snapshot (`CC-MAIN-2026-04`, 396,347 records, ∼12 GB) has completed successfully; the remaining snapshots are queued for incremental processing. The pipeline can be run continuously as new snapshots are published, growing the corpus over time without re-processing earlier data.

1. **Manifest generation (index level).** Either via Athena SQL or the Common Crawl Index API. Typical filters include `status=200`, `mime=text/html`, and `language=sqi`. The output is a CSV with WARC filename, byte offset, record length, and metadata.

2. **Reproducible download.** Records are retrieved by HTTP range requests against Common Crawl's S3-hosted WARC files. Neighboring byte ranges are merged to reduce the number of requests. Each attempt is logged in the fetch ledger (success/failure, HTTP status, content hash, timestamp), making runs resumable: on restart, already-fetched records are skipped. From the fetch ledger, a success manifest of retrieved records can be exported and shared so that others can retrieve the same record set.

3. **Cleaning and language verification.** Text is extracted from the retrieved WARC payloads. Heuristic quality checks (minimum text length, alphabetic character ratio, token repetition, boilerplate ratio) and a language-specific plausibility check are applied; each decision is logged to a separate cleaning ledger. The language check and the subsequent FastText filtering stage are described in Section 5.1.

4. **Deduplication and perplexity scoring.** Near-duplicate detection via shingling (Broder, 1997) removes redundant content across records. Documents that pass deduplication are optionally scored for perplexity using a KenLM n-gram language model (Heafield, 2011); implausibly high perplexity (indicating garbled text, encoding artifacts, or non-natural-language content) is a drop signal.

5. **Learned filtering.** A larger LLM labels a stratified sample of cleaned documents for quality; these labels are used to train or calibrate a faster student filter. The student's keep/drop decisions are again logged per record and exported as a keep manifest. This stage is described in detail in Section 6.

6. **Reporting.** A summary of the full pipeline funnel—records in, records kept at each stage, failure categories, and deduplication statistics—is generated from the ledgers.

The release artifact is therefore the pipeline itself rather than a text dump.

5.1. Language Filtering for Albanian

Language filtering proceeds in two stages, each targeting a different failure mode.

Heuristic pre-filter. The cleaning step includes a lightweight Albanian plausibility check that serves as a fast pre-filter before the more expensive FastText classification. It combines two signals: (a) the ratio of Albanian stopwords among all tokens, and (b) the ratio of Albanian-specific diacritics—specifically *ë* and *ç*—to total characters. These signals are combined into a composite score in

which the diacritic ratio is weighted substantially higher than the stopword ratio (currently by a factor of 12, chosen empirically). The rationale is linguistic: while Albanian shares many high-frequency function words with neighboring Balkan languages, the diacritics *ë* and *ç* are distinctive and occur frequently in Albanian text but rarely in Romanian, Serbian, or other regional languages. This makes the diacritic signal a strong discriminator even for short documents where stopword counts alone would be unreliable. The exact weighting factor is a candidate for systematic ablation; as part of our evaluation, we also test the pipeline without the heuristic pre-filter entirely, relying on FastText alone, to measure its contribution to precision and throughput.

FastText verification. Documents that pass the heuristic check are classified by a FastText language-identification model (Joulin et al., 2017). Rather than a single confidence threshold, the filter applies a two-tier acceptance strategy: a document is kept if the model's top prediction is Albanian with confidence ≥ 0.80, *or* if Albanian appears among the top three predictions with confidence ≥ 0.60. These thresholds are conservative defaults; the ledger-based design makes it straightforward to re-evaluate with different values without re-downloading or re-extracting text. The second, more permissive gate addresses a pattern common in Albanian web content: news articles and institutional pages frequently contain English headlines, embedded quotations in other languages, or code-switched passages that depress the top-1 confidence for Albanian while still leaving it among the most likely languages. Without this fallback, such documents—which are often substantively Albanian—would be systematically lost.

Both stages log their decisions per record (including the scores, thresholds applied, and reason codes), so the effect of each gate can be inspected and adjusted in subsequent runs.

To illustrate the interaction of these filters: a news article from a `.al` domain with consistent Albanian prose, frequent use of *ë*, and a FastText top-1 confidence of 0.92 passes both stages easily. An Albanian government page with an English-language header and bilingual body might score only 0.65 on top-1 but 0.72 on top-3, and would be rescued by the fallback gate. Conversely, a page in Macedonian that was misclassified as `sqi` in the Common Crawl index would fail both the diacritic check (Macedonian uses Cyrillic or Latin without *ë*) and the FastText gate. A cookie-consent dialog or navigation-only page, even if in Albanian, would be caught earlier by the minimum-text-length and boilerplate-ratio checks in the cleaning stage.

5.2. Operational Properties

Index acquisition and download are resumable: the builder persists state per page and per TLD, so interrupted runs can continue without re-fetching. Rate limiting from Common Crawl's servers is handled by circuit-breaker logic with cooldown windows; when a TLD's index queries are repeatedly throttled, that TLD is temporarily set aside and retried later. When records from multiple crawl snapshots are merged, explicit priority rules and deterministic tie-breaking ensure reproducible deduplication outcomes.

6. Learned Filtering Stage

The learned-filter stage applies a score-based keep/drop policy in two steps:

- `evaluate-threshold` selects a score threshold from labeled data under precision/recall constraints (e.g., "keep at least 90% of good documents while dropping at least 70% of noise"),
- `apply` applies this threshold to all documents: each document's score is compared against the threshold, and the keep/drop decision is logged with a machine-readable reason.

The stage is independent of how scores are produced. In the teacher–student setup, scores come from a student model trained on LLM-generated labels; the same policy mechanism works with any external scorer. The end-to-end quality gain of this stage over baseline filtering is not yet demonstrated on the final corpus.

6.1. Segment-First Student Filtering Architecture

We define the student stack as a segment-first cascade with four deterministic stages:

1. **Stage S (BlockTagger Student).** Segment-aware classification predicts functional block labels (for example MAIN/TITLE vs NAV/COOKIE/FOOTER/ADS/TEMPLATE). Outputs are (a) canonical main text, (b) boilerplate ratios, and (c) template signatures.
2. **Stage Q (DocScorer Student).** A document-level scorer produces multiple output signals from main text and structural features: a quality score, a language-purity estimate, a content-type label, and a preliminary drop reason. These signals are computed jointly by a single model with multiple output heads.
3. **Stage R (Compact Reranker, optional).** Documents where Stage Q's quality score falls in an uncertain range (neither clearly good nor

clearly bad) are re-scored by a small transformer model. The majority of documents bypass this stage entirely, keeping throughput high.

4. **Stage D (Deterministic set selection).** Final keep decisions are computed under cluster constraints (near-duplicate clusters on main text and template clusters on boilerplate signatures), without per-host caps.

This decomposition allows each stage to be evaluated and ablated independently.

6.2. Training and Policy Learning

Training follows a teacher-to-student distillation approach. A large open-weight LLM serves as the teacher: we currently use GPT-OSS-120B, a 120-billion-parameter model that can be run locally on desktop hardware with large unified memory, such as NVIDIA DGX Spark or AMD Ryzen AI MAX 395 systems with 128 GB LPDDR5x. As a comparison teacher we are evaluating Qwen3.5-27B on workstations with NVIDIA RTX 5090 GPUs. Inference is served locally via llama.cpp and LM Studio. While this hardware is not inexpensive, it is far below the scale of a data-center deployment: a small number of local machines can label stratified samples at sufficient throughput for a single-language corpus.

The distillation process works as follows:

1. **Teacher labeling.** The teacher LLM labels stratified samples of cleaned documents at document level, assessing content quality, language purity, and topical relevance on a structured rubric. Stratification ensures coverage across quality tiers, domains, and document lengths so that the student sees representative positives and negatives. Where segment-level annotation is available, the teacher additionally labels functional blocks (main content vs. boilerplate).

2. **Student training.** Segment-level labels train the BlockTagger (Stage S); document-level quality and preference-style targets train the DocScorer (Stage Q). The optional reranker (Stage R) is trained only on cases where Stage Q is uncertain.

3. **Threshold calibration.** Deployment thresholds are selected on held-out labeled data under precision/recall constraints (e.g., keeping $\geq$90% of teacher-approved documents) and then frozen for the production run.

4. **Deterministic execution.** Stage D applies fixed tie-break rules and emits machine-readable drop reasons (e.g., near-duplicate, template-dominated, mixed-language, boilerplate-heavy).

For Albanian, training and evaluation emphasize hard negatives that are characteristic of the language's web presence: code-switching with closely related Balkan languages, template-heavy pages from a small number of dominant domains, and thin SEO or listing content.

7. Conclusion and Future Work

Merimënga demonstrates that reproducible web-corpus construction for a single language does not require the infrastructure of a large multilingual project. By querying Common Crawl's index for Albanian records and fetching only those via byte-level range requests, the pipeline reduces the data volume by orders of magnitude, making it feasible to run on desktop hardware. The construction process—from index query to final keep/drop decision—is fully recorded and can be replayed, inspected, and extended by third parties.

By separating the reusable construction recipe from the derived text, the manifest-first release model reduces redistribution concerns while preserving transparency. New crawl snapshots can be incorporated incrementally, and filter parameters can be revised without re-downloading data. For lower-resource languages, where research groups cannot always wait for the next release of a multilingual project, this independence is a practical advantage.

Planned evaluation work includes systematic ablation of the filter stack (heuristic pre-filter, FastText thresholds, learned filter), measurement of rerun equivalence across independent pipeline executions, and reporting of per-stage yield and drop-reason distributions once the full multi-snapshot run completes.

8. Ethics Statement

Web-crawled text can contain copyrighted material, personal data, and sensitive content. To reduce redistribution risk, we release manifests and code rather than full text; users retrieve content directly from Common Crawl. Each filtering decision is logged with a machine-readable reason, enabling post-hoc inspection. The approach does not guarantee that reconstructed material is free of personal or sensitive information; downstream users remain responsible for additional safeguards appropriate to their jurisdiction and use case.

9. Limitations

Full-corpus filtering is ongoing; we do not yet report final corpus size, yield, or end-to-end quality comparisons. Reproducibility depends on continued public availability of the referenced Common Crawl

snapshots; changes in access patterns or snapshot retention can affect replay feasibility. Deterministic record identifiers provide stable addressing within a snapshot but do not eliminate all sources of variance (e.g., transient retrieval failures or extraction corner cases), which are mitigated operationally via resumable ledgers and integrity checks.

Language identification and quality filtering are imperfect for Albanian, particularly in the presence of code-switching with closely related Balkan languages, template-heavy pages, and domain concentration. Learned filtering inherits the assumptions of the teacher model and may amplify annotation biases; we do not yet claim validated quality improvements from this stage on the final corpus.

10. Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoit Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Julien Abadji, Pedro Javier Ortiz Suarez, Laurent Romary, and Benoit Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9)*, pages 1–9, Mannheim. Leibniz-Institut fur Deutsche Sprache.

Manuel Brack, Malte Ostendorff, Pedro Ortiz Suarez, Jose Javier Saiz, Inaki Lacunza Castilla, Jorge Palomar-Giner, Alexander Shvets, Patrick Schramowski, Georg Rehm, Marta Villegas, and Kristian Kersting. 2024. Community oscar: A community effort for multilingual web data. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 232–235, Miami, Florida, USA. Association for Computational Linguistics.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29.

Common Crawl Foundation. 2024. Common crawl terms of use. https://commoncrawl.org/terms-of-use. Accessed 2026-02-25.

Common Crawl Foundation. 2026a. Common crawl: Get started. https://commoncrawl.org/get-started. Accessed 2026-02-25.

Common Crawl Foundation. 2026b. Common crawl index. https://index.commoncrawl.org/. Accessed 2026-02-25.

Common Crawl Foundation. 2026c. Common crawl on aws open data registry. https://registry.opendata.aws/commoncrawl/. Accessed 2026-02-25.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Hugging Face. 2024. Huggingfacefw/fineweb-edu-classifier - model card. https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier. Accessed 2026-02-25.

International Organization for Standardization. 2017. Iso 28500:2017 information and documentation – warc file format. International Standard.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 427–431.

Jeffrey Li, Alex Fang, Georgios Smyrnis, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9–16, Cardiff, UK. Leibniz-Institut fur Deutsche Sprache.

OSCAR Project. 2019. Oscar 2019 - oscar documentation. https://oscar-project.github.io/documentation/versions/oscar-2019/. Accessed 2026-02-25.

OSCAR Project. 2021. Oscar 21.09 - oscar documentation. https://oscar-project.github.io/documentation/versions/oscar-2109/. Accessed 2026-02-25.

OSCAR Project. 2022. Oscar 22.01 - oscar documentation. https://oscar-project.github.io/documentation/versions/oscar-2201/. Accessed 2026-02-25.

OSCAR Project. 2023. Oscar 23.01 - oscar documentation. https://oscar-project.github.io/documentation/versions/oscar-2301/. Accessed 2026-02-25.

Guilherme Penedo, Hynek Kydlicek, and all. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Guilherme Penedo, Hynek Kydlicek, Loubna Ben Allal, et al. 2024. Fineweb: Decanting the web for the finest text data at scale.

David Samuel, Andrey Kutuzov, Satya Almasian, et al. 2024. Hplt: High-performance language technologies datasets.

Roland Schäfer and Felix Bildhauer. 2016. Common crawl and web corpus construction: Practical and legal considerations. https://commoncrawl.org/blog. Accessed 2026-02-25.

Webrecorder. 2026. warcio: Streaming warc reader/writer for python. https://github.com/webrecorder/warcio. Accessed 2026-02-25.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzman, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Pop Lyrics Through Time: Challenges in Corpus-Based Modeling of Linguistic and Emotional Dynamics in German Pop Lyrics

Roman Schneider

Leibniz Institute for the German Language (IDS)
R5 6-13, 68161 Mannheim / Germany
schneider@ids-mannheim.de

Abstract

This paper presents a large-scale diachronic analysis of German pop lyrics based on a linguistically rich, TEI-encoded monitoring corpus. We describe multi-layer annotation and reproducible workflows for deriving higher-level features at scale, including lexical diversity indices, a pronoun-based subjectivity measure, modal particle density, and a length-normalized sentiment intensity score. Particular attention is paid to the development and evaluation of pipelines for two notoriously challenging phenomena: modal particles and sentiment. For modal particles, we build a manually curated gold standard and train sequence models whose performance we relate to inter-annotator agreement. For sentiment, we integrate a lexicon-based resource with a dedicated human annotation experiment to assess reliability and alignment with expert judgments. On this basis, we investigate how structural and affective features co-vary in the corpus and how they change over time, showing, among other trends, declining lexical diversity and sentiment intensity alongside a slight increase in first- and second-person pronouns. Beyond the empirical findings, the paper highlights practical challenges in managing culturally specific corpora, and makes evaluation materials available to support transparent, reusable corpus-based research on popular music and related domains.

Keywords: culture-specific corpora; diachronic analysis; annotation quality

1. Introduction

The empirical study of language increasingly depends on large-scale resources that capture diverse communicative domains. While substantial progress has been made in developing corpora for functional, non-literary registers, popular culture remains comparatively underrepresented. Pop song lyrics – despite their considerable communicative "impact factor" (Kreyer and Mukherjee, 2007) – have traditionally been relegated to literary and cultural studies rather than linguistics, reflecting both disciplinary conventions and the former absence of sustainable reference corpora.

Recent years have seen a shift, as research on lyrics has begun to gain visibility within empirical linguistics (cf. Werner et al., 2025), supported by the emergence of dedicated resources (Schneider, 2026). For German, the release and continuous expansion of the extensively annotated Songkorpus (Schneider, 2020, 2025) has established pop lyrics as a legitimate and sustainable object of quantitative investigation. From a linguistic perspective, sung – or in some cases more spoken, as in rap lyrics are particularly interesting because they constitute a hybrid text type that draws on features of both conceptual orality and literacy. Songs often incorporate colloquial and spoken-language elements while at the same time adhering to genre-specific structural and stylistic conventions. This duality makes lyrics a valuable source for studying contemporary language use, the interaction between spoken and written registers, and broader cultural dynamics.

Building on work that locates texts in a hybrid space between "language of immediacy" and "language of distance" (Koch and Oesterreicher, 2012), we combine structural features with sentiment analysis to provide a multidimensional account that integrates both linguistic and affective dimensions. At the word-form level, features such as modal particle proportion, pronoun proportion, and lexical richness have been shown to be key predictors of a text's position in this hybrid space (Broll and Schneider, 2023). Our approach extends this by adding an empirically validated measure of sentiment intensity, enabling us to examine how markers of proximity and distance interact with affective tone.

On the empirical side, large-scale studies of English lyrics report that lyrics have become simpler and more repetitive in terms of lexical richness, readability, and repetition over the last five decades, with rap retaining relatively high complexity (Parada-Cabaleiro et al., 2024). For German popular music, Hunke et al. (2025) find increasingly negative tendencies in chart songs, though their claims are constrained by multilingual sampling and translation into English. This paper addresses both lexical and affective developments in a substantially larger, German-only corpus and explicitly links them to core features of German grammar and discourse such as modal particles and pronouns.

The selection of linguistic features in this study is guided by the assumption that song lyrics are not only vehicles of thematic content, but also sites of perspective-taking, stance marking, and affective expression. Pronoun usage provides a window into how subjectivity and interpersonal orientation

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 32–43. Palma, Mallorca. May 11, 2026. ©ELRA Language Resources Association, 2026

are linguistically constructed, for instance through shifts between self-reference and addressivity. Modal particles, a characteristic feature of German interactional language, index speaker stance, pragmatic nuance, and degrees of communicative immediacy. Sentiment measures, in turn, offer a coarse-grained approximation of affective framing at scale. Taken together, these features are intended to capture complementary dimensions of how lyrics position speakers, encode attitudes, and shape emotional expression in a genre that is simultaneously written, performed, and socially embedded.

This paper is conceived as a methodological contribution to corpus-based research on creative, non-standard language. Its central aim is to demonstrate robust strategies for extracting linguistic and affective features from song lyrics, a domain that systematically challenges standard NLP assumptions. The substantive analysis of German pop lyrics over time serves as an empirical test case that illustrates the potential and limitations of these methods. While the findings offer insights into linguistic and emotional dynamics, they are intended as exploratory and methodologically grounded rather than as definitive claims about cultural change.

1.1 Related work

In the broader domain of song lyrics studies, Parada-Cabaleiro et al. (2024) analyze 12,000 English-language songs across five genres, focusing on lexical diversity, readability, and repetition. They report that lyrics have become simpler and more repetitive over time, with lexical richness declining and repetition increasing. Rap stands out as retaining higher lexical complexity relative to other genres. These findings raise the question of whether similar trends can be observed in other languages and corpora with different sampling strategies.

Modal particles (MPs) are a quintessentially German word class, used to signal subjectivity, discourse stance, or speaker-listener intimacy. Their insertion or omission can subtly shift force, politeness, or hedging. Because MPs sit at the interface of semantics, pragmatics, and discourse, they resist simple categorization: they overlap with adverbs, discourse particles, or focus markers, and their acceptability is context- and order-sensitive (Diewald, 2007; Blühdorn, 2019; Schoonjans, 2018). Frequent homography and context dependence complicate automated detection (Storø, 2022), and many standard tag sets do not label MPs separately. While individual studies have examined MPs in literary genres (e.g. Hentschel, 2010), large-scale analyses on pop corpora are lacking.

Sentiment analysis of song lyrics has begun to attract attention. Hunke et al. (2025) compute topic and sentiment models for German chart songs and report increasingly negative tendencies over time, though their sample is relatively small and multilingual. Beyond bag-of-words scoring, discourse-aware models such as Discourse-LSTM (Kraus and Feuerriegel, 2019) incorporate rhetorical structure to mitigate position biases. From the broader NLP literature, classic overviews (Pang and Lee, 2008; Liu, 2015) emphasize pitfalls that are especially relevant for lyrics: irony, context dependence, lexicon domain mismatch, polysemy, and short-text limitations. At the same time, contemporary sentiment research increasingly employs neural and transformer-based models that integrate context and negation, though their behavior on poetic and lyrics data has not yet been systematically studied.

Across all phenomena under investigation, the Tool Misuse perspective (Sluyter-Gäthje and Trilcke, 2022) is relevant. It reminds us that the “errors” of NLP tools on literary texts may reflect systematic stylistic deviation, while such tools remain essential for analyzing large corpora.

1.2 Research questions and contribution

We address three research questions that bridge methodological and substantive concerns:

- **RQ1:** How reliably can automated methods detect notoriously difficult-to-capture linguistic and affective phenomena – specifically modal particles and sentiment intensity – as assessed against manually curated gold standards or experiments involving human participants?
- **RQ2:** To what extent do lexical-syntactic and emotional phenomena co-vary? In particular, how are lexical diversity, modal particle usage, pronoun use, and sentiment intensity associated, as evaluated using correlation and regression analyses?
- **RQ3:** Can temporal patterns or trends be identified across the dataset, indicating systematic change over time in these structural and affective features?

Taken together, these questions are designed to first establish the reliability of the analytical approach (RQ1), and then to explore its application to diachronic linguistic and affective patterns (RQ2, RQ3).

The study’s contributions are fourfold:

1. We use Songkorpus, a large, multi-layer annotated corpus of German song lyrics, to conduct a diachronic analysis spanning more than five decades.
2. We develop and evaluate an automated approach to modal particle identification in lyrics, grounded in a manually annotated gold standard.

3. We integrate sentiment analysis with a dedicated human annotation experiment to derive and validate a measure of sentiment intensity at the song level.
4. We model the interrelations and temporal development of lexical diversity, modal particles, pronouns, and sentiment intensity, thereby linking structural features of German with affective expression in popular music.

In Section 2, we address RQ1 by detailing the corpus, feature extraction, and validation of the most error-prone automatic annotations, before turning in Section 3 to RQ2 and RQ3, where we analyze interrelations among the features and their diachronic development.

2. Data and methods

2.1 Corpus

Songkorpus (Schneider, 2025) is currently the largest scientifically curated monitoring collection of German-language song lyrics and constitutes a unique public resource for linguistic and cultural research. With more than 15,000 lyrics and approximately five million tokens, it covers over six decades of music history and continues to be updated. Archives are revised annually (e.g. through the addition of current chart songs), and recent developments include an expanded hip-hop ("Deutschrap") archive reflecting the official German hip-hop charts.

The corpus is systematically stratified into multiple archives. Artist-specific archives collect the complete works of established performers and emerging artists. Thematic archives are organized along historical lines – such as songs from the former GDR – or by genre, including a dedicated archive of early-1980s "Neue Deutsche Welle". A monitoring archive covers all German-language chart songs since the mid-1950s; for this study, however, we focus on data from 1970 onward, as this reflects the more data-intensive period, enabling robust temporal stratification and meaningful diachronic analysis. This combination allows the data set to represent both mainstream and subcultural repertoires across a wide range of themes and time periods.

Intellectual property considerations play a central role in the compilation and maintenance of the Songkorpus. As song lyrics are typically protected by copyright, full-text data cannot be freely redistributed. The corpus therefore follows a controlled access model: while the texts are stored and processed for research purposes, public distribution is restricted to derived data (e.g. annotations, frequency information, or aggregated features). For artist-specific archives, this framework is complemented by individual agreements with rights holders, which allow more extensive use under clearly defined conditions.

Each text is encoded in TEI format and enriched with bibliographic metadata (e.g., author, release year, genre, source); the central importance of metadata for linguistic data is discussed in Trippel (2025). Crucially, lyrics are annotated at multiple linguistic levels, including lemmata, part-of-speech tags (STTS), named entities, neologisms, syntactic constituents, verse structures, and, in some cases, rhyme schemes. The corpus is accessible via an online portal with tailored search forms, visualizations, and downloadable datasets (bag-of-words, n-grams, word vectors).

To manage the corpus as a continuously updated monitoring resource, we rely on an automated processing pipeline that standardizes ingestion, annotation, and export of derived variables (cf. Schneider, in preparation). Crucially, this pipeline also computes the linguistic and affective features that we investigate in the present study.

Code-switching between German, English, and other languages is a frequent feature of contemporary song lyrics and is explicitly represented in the Songkorpus. The corpus uses TEI-based language attributes to mark segments at a fine-grained level. This allows non-German material (e.g. English insertions or multilingual passages) to be systematically identified and filtered. For the present analyses, only segments annotated as German are included to ensure methodological consistency.

2.2 Feature set and scope

We focus on four feature families that jointly capture aspects of lexical richness, interpersonal alignment, discourse marking, and affective density:

- **Lexical diversity** (MATTR, MTLD) as indicators of lexical richness and, indirectly, of how planned and elaborated (more distant) vs. formulaic and repetitive (more immediate) the language is.
- **Pronoun index** (PRON) to capture subjectivity and address patterns through first-/second- vs. third-person forms.
- **Modal particle ratio** (PTKM) as a core marker of German discourse stance, subjectivity, and speaker–listener intimacy.
- **Sentiment intensity** (SI) as a length-normalized measure of affective density, irrespective of polarity.

This selection does not exhaust the space of relevant features. We do not explicitly model syntactic complexity, rhythmic structure, or topic/lexical field distributions, nor do we include discourse-structural sentiment models or multi-dimensional emotion categories. We therefore understand our account as *multidimensional* but not fully comprehensive: it targets key lexical-syntactic and affective indicators while leaving other dimensions to future work.

2.3 Linguistic features

2.3.1 Lexical diversity

Lexical diversity is a key criterion in distinguishing between spoken and written language, or more generally between language of immediacy and language of distance (Malvern et al., 2004; Koch and Oesterreicher, 2012). Simple type-token ratios (TTR) are strongly correlated with text length and thus limited in interpretive value. A Pearson correlation across the corpus reveals a moderate positive relationship between token count and year of publication ($r \approx 0.40$), indicating that song texts tend to become longer over time; the number of texts per year also increases slightly. Both trends motivate the use of more robust metrics.

We consider three established measures:

- **Standardized TTR (STTR)**, which divides texts into successive windows of fixed size (here: 100 tokens) and computes the mean TTR across segments (Scott, 2004).
- **Moving-Average TTR (MATTR)**, which applies a sliding window over the text and averages TTR across overlapping windows, thereby avoiding partial final segments (Covington and McFall, 2010).
- **Measure of Textual Lexical Diversity (MTLD)**, which estimates diversity through sequential passes (forward and backward) and is designed to be comparatively robust to length (McCarthy and Jarvis, 2010).

All three measures are sensitive to very short texts, a general limitation emphasized by Bestgen (2024). In song lyrics, this is particularly relevant because many songs would be relatively short without repeated refrains and hooks, which at the same time introduce strong local repetition. After initial exploration of STTR, MATTR, and MTLD, we retain the latter two: MATTR is particularly attractive for this genre because its sliding-window procedure is sensitive to the recurring, stylistically driven repetitions typical of musical texts. MTLD, by contrast, assesses lexical diversity sequentially across the entire text and captures how repetitions influence lexical variation at a more global level. Using both measures thus allows us to compare a locally sensitive, window-based approach with a sequential, text-wide measure, and to evaluate which metric more adequately handles the specific properties of pop lyrics, also in interaction with the other features.

In this study, we compute both measures for all texts, report their correlation, and explore their differential behavior. In the subsequent regression models, we focus on MATTR, which yields more stable and interpretable associations with other features in this corpus, and we explicitly revisit the divergence between MATTR and MTLD in the discussion.

2.3.2 Pronoun index

Personal and possessive pronouns are included as features because they index interpersonal relations and subjectivity. First- and second-person pronouns (German: *ich, du, mein, dein*, etc.) are comparatively infrequent in conventional written corpora, where third-person forms (German: *er, sie, es, sein, ihr,* etc.) predominate, but are expected to occur more frequently in song lyrics, likely reflecting their orientation toward intimacy, immediacy, and emotional expression. The lyrical subject ('I') and addressee ('you') are frequently positioned in relation to, or in contrast with, a third person ('he/she'). We operationalize this by measuring the polarity between first- and second-person pronouns versus third-person pronouns using automated part-of-speech tagging combined with curated wordform lists. As with the lexical diversity measures, we compute a single score (PRON) for each text.

2.3.3 Modal particles

Modal particles pose a particular challenge in corpus-based research due to their polyfunctionality and homonymy with other word classes. Identification based solely on word forms is error-prone, as many candidate forms also occur as adverbs, connectors, or discourse markers. While MPs in German typically appear in the Mittelfeld position, their placement is flexible and influenced by discourse and syntax, and their subtle pragmatic functions cannot be captured by syntax alone.

To illustrate the linguistic behavior of MPs, consider a few examples. **„**Aber**“** can convey emphasis or surprise: *„Das ist aber schön geworden!“* ["That has really turned out nicely!"] signals positive astonishment, while *„Du bist aber spät dran“* ["You are really late"] indicates mild criticism. **„**Doch**“** often marks known or expected information, e.g., *„Das ist doch nicht so schlimm“* ["That's really not so bad"], or introduces contrast, as in *„Er wollte zuerst nicht, doch er kam trotzdem“* ["He didn't want to at first, but he came anyway"]. **„**Bloß**“** can relativize or warn: *„Mach bloß keinen Blödsinn!“* ["Just don't do anything stupid!"] highlights caution, whereas *„Er hat bloß Pech gehabt“* ["He just had bad luck"] downplays the situation. Other particles include **„**ja**“** (*„Du weißt das ja auch“* ["You know that too"]) for self-evident information, **„**mal**“** (*„Ich schau mal, ob ich Zeit habe“* ["I'll just see if I have time"]) for tentative or polite action, and **„**vielleicht**“** (*„Ich war vielleicht überrascht“* ["I was maybe surprised"]) for uncertainty or hedging. These examples show how modal particles operate in subtle, context-dependent ways, reinforcing the need for context-sensitive modeling.

We target a core inventory of 14 high-frequency MPs: *aber, auch, bloß, denn, doch, eben, etwa, halt, ja, mal, nur, schon, vielleicht*, and *wohl*. Without making any definitional claim, this set is widely recognized as constituting a stable MP core. We adopt a two-step approach (Section 2.5): (i) create a manually annotated gold standard for these forms in context, then (ii) train and evaluate a sequence model for automatic MP detection. For the corpus-level analyses, we compute PTKM, the ratio of automatically detected MPs to total tokens per text.

2.4 Sentiment and emotional modeling

2.4.1 Automated sentiment computation

Most classical sentiment approaches in NLP rely on lexicons and rule-based heuristics; more recent work increasingly uses neural and transformer-based models that integrate contextual information. For German lyrics, however, large genre-specific labeled datasets for supervised training are scarce, and interpretability is a concern. We therefore adopt a lexicon-based approach as a transparent and easily interpretable baseline.

Specifically, we use SentiMerge (Emerson and Declerck, 2014), developed to integrate and harmonize multiple German sentiment resources. SentiMerge assigns sentiment scores to lemmata based on several lexica and applies weighting schemes that account for both reliability and frequency. For example, the adjective *abscheulich* 'abhorrent' receives a strongly negative score of approximately -0.9 with a weight of 9.7, while the noun *Abgott* 'idol' is assigned a positive score of +0.9 with a weight of 6.7.

To apply SentiMerge consistently, we harmonize POS tags between the STTS tagset used in Songkorpus and SentiMerge's categories: all nouns are mapped to "N", full verbs in all forms to "V", and other categories (e.g. onomatopoetic forms, discourse markers) to the closest available types. Sentiment is computed at the level of the entire song text, which we treat as a coherent unit, consistent with the other features. While verses or stanzas may vary in tone, the present study focuses on overall song-level affect rather than within-song dynamics.

Simple sentiment sums depend on text length: longer songs naturally accumulate more sentiment-bearing words, regardless of orientation. Preliminary analyses show that the songs with the highest total positive sentiment also rank highest in total negative sentiment, and total sentiment is negatively correlated with token count. We therefore experiment with several normalizations, including division by total tokens and by sentiment-bearing tokens. Polarity measures (ratio of positive to negative sums) turn out to be relatively stable over time, showing only a slight shift toward negativity and providing limited diachronic discrimination.

For our main analyses, we focus on Sentiment Intensity (SI), defined as the sum of the weighted absolute values of all sentiment scores in a song divided by token count. SI captures affective density (how strongly emotional language is mobilized) without privileging positive or negative orientation. This aligns with our primary interest in *how much* emotional language is used, rather than in assigning songs an overall positive or negative label or distinguishing between specific evaluative stances.

2.4.2 Why not transformer-based sentiment?

Transformer-based models offer powerful context-sensitive sentiment and emotion classification and can handle negation and long-distance dependencies more effectively than lexicons. However, there are several reasons why we do not use them here:

1. **Domain mismatch and training data.** Existing German BERT-based sentiment models are primarily trained on user-generated content such as tweets, online posts, and product or movie reviews, rather than on lyrics or poetry (Bello et al, 2023; Guhr et al., 2020). Across languages, work that targets song lyric relies on transfer learning from other domains: for instance, emotion models for pop lyrics pre-train on large generic or social-media corpora and are then fine-tuned on comparatively small lyric datasets (Dahary et al, 2025). To our knowledge, there are currently no widely used German transformer-based sentiment models that are trained directly on lyrics, and we are not aware of systematic evaluations of such models. Where literary texts are analyzed, authors typically highlight data scarcity and therefore rely on cross-domain transfer or on lexicon-based methods adapted to the target domain (Öhman, 2021; Fehle et al., 2021).
2. **Interpretability.** Our aim is to relate sentiment to lexical-syntactic features. Lexicon-based scores make this mapping explicit; neural models are less transparent.
3. **Cross-temporal and cross-linguistic comparability.** Lexicon-based approaches can more easily be harmonized across decades and languages in future work, whereas neural models may introduce additional diachronic and domain-specific biases.
4. **Expected gains relative to task difficulty.** As shown in Section 2.5.2, even human annotators exhibit only moderate agreement at the aggregate level, with most disagreements involving adjacent categories rather than polarity reversals. This suggests that

sentiment in song lyrics is often inherently ambiguous or underspecified. In such settings, more complex models may yield only limited improvements over simpler approaches.

We therefore treat SentiMerge as a robust, interpretable baseline for song-level affect. At the same time, the gold-standard annotations created in this study provide a useful testbed for future work: transformer-based classifiers or prompting-based large language models could be evaluated systematically against human judgments to assess whether they yield measurable improvements or different patterns of disagreement. Future work, as discussed in the conclusion, should test whether such approaches produce different diachronic patterns or stronger associations with structural features.

2.5 Evaluation of automatic annotation (RQ1)

RQ1 concerns the quality and reliability of the automatic methods used to derive our most error-prone key features, namely modal particles and sentiment intensity. We therefore conduct two dedicated evaluations. Other features (lexical diversity measures and pronoun counts) are based on relatively robust, well-studied procedures (tokenization, lemmatization, POS tagging, dictionary lookup) and are therefore used descriptively without additional task-specific validation in this study.

2.5.1 Modal particle gold standard and CRF models

We construct a gold standard based on a stratified sample of 1,400 instances of the 14 target word forms (100 per form), drawn from Songkorpus. Each example is presented with surrounding context (typically one sentence to the left and right), and four native speakers independently annotate whether the candidate functions as an MP or not, following detailed guidelines with positive and negative examples.

Inter-annotator agreement is substantial to almost perfect (Landis and Koch, 1977): pairwise Fleiss' κ values range from 0.64 to 0.83, with most above 0.75. Certain particles present systematic challenges; *mal*, for example, is highly polyfunctional (emphasis, softening, temporal adverbial uses), and the lack of prosodic cues in written data increases variability. Disagreements are resolved by majority vote; ties are adjudicated by a trained annotator, yielding a curated gold standard.

Building on this resource, we train Conditional Random Field (CRF) models using CRF++ (Kudo, 2005) for token-level MP classification (MP vs. non-MP). Features include token form, lemma, and STTS tags within variable context windows. We compare three configurations:

- Reduced model (narrow context ±2, basic lemma and POS features).
- Original CRF++ model (same context, richer feature conjunctions).
- Extended model (broader context ±3).

Contrary to expectations, the reduced model performs best, correctly identifying 202 of 286 MPs in the test set and achieving high precision and recall for particles such as *etwa*, *halt*, and *wohl* (>90%). Performance for more polyfunctional items (*ja, vielleicht, schon*) is weaker, mirroring human disagreement patterns. A simple wordform baseline that classifies all occurrences of the 14 forms as MPs would perform reasonably for some particles but poorly overall, underscoring the benefit of context-sensitive modeling.

The close alignment between human variability and model performance suggests that the limitations of automated MP identification partly reflect ambiguity inherent in the data. Overall, the reduced CRF model provides sufficient quality for corpus-level modeling, with known weaknesses explicitly acknowledged.

2.5.2 Sentiment annotation experiment

To assess the fundamental quality of SentiMerge-derived sentiment scores, we conduct an annotation experiment that compares system scores with human judgments and tests intra-rater stability (Abercrombie et al., 2023).

We select 20 songs: five that SentiMerge classifies as very positive, five as very negative, five with moderately positive scores, and five with moderately negative scores (based on corpus-level means). Seven raters classify each song twice, separated by several days, using four ordered categories: very negative, rather negative, rather positive, very positive. This yields 280 decisions (20 songs × 7 raters × 2 sessions).

We encode the categories on an ordinal scale that treats differences as distances: one point for adjacent categories, two points when one category lies in between, and three for extremes, positive versus negative. This allows us to quantify both categorical and graded disagreements.

Three complementary evaluations are conducted:

1. **Inter-rater reliability.** Krippendorff's α across raters is around 0.30 for both sessions, indicating modest aggregate agreement. Average pairwise weighted Cohen's kappa κ is substantially higher (≈0.71 – 0.72), suggesting that most rater pairs are consistent and that the low α reflects outliers.
2. **System-human alignment.** Weighted κ between SentiMerge and individual raters ranges from ≈0.52 to 0.76 (moderate to

substantial agreement), with accuracy scores between 0.5 and 0.7 and Mean Absolute Error (MAE) between 0.5 and 0.8. Disagreements mainly involve neighboring categories rather than opposites, indicating that the system rarely mistakes strongly positive for strongly negative songs or vice versa.

3. **Intra-rater reliability.** Weighted kappa between each rater's two sessions is very high (0.92 – 1.00), with most values ≥0.96, indicating that individual judgments are highly stable over time.

Taken together, these results confirm that SentiMerge-based scores operate within the same reliability range as human judgments for the selected subset. They provide a justified basis for using sentiment intensity as a corpus-level measure, while acknowledging that subtle nuances, ironic uses, or complex metaphors remain challenging.

3. Empirical Results

Having established the robustness and limitations of the feature extraction procedures, the now following section applies these methods to the diachronic analysis of German pop lyrics.

Given that Section 2 offers a largely positive answer to RQ1 for modal particles and sentiment intensity, we now turn to RQ2 and RQ3. First, we examine how the four feature families co-vary (correlations and regression models), then we model their temporal development.

3.1 Associations between structural and affective features (RQ2)

We compute Spearman's ρ correlations among lexical diversity (MATTR, MTLD), sentiment intensity (SI), modal particle ratio (PTKM), and pronoun index (PRON). The resulting matrix shows the following patterns:

- **Lexical diversity.** MATTR and MTLD are only weakly correlated (≈0.07), suggesting that they capture different aspects of diversity in this corpus. This may reflect the fact that MATTR is sensitive to local repetition (e.g. choruses), whereas MTLD is driven by global variation.
- **Sentiment intensity.** SI correlates strongly and positively with MATTR (≈0.70), suggesting that lexically more diverse songs more densely mobilize sentiment-bearing words. SI also shows a moderate positive association with PTKM (≈0.29), implying that MP-rich texts tend to be more affectively loaded.
- **Modal particles.** MATTR shows a moderate positive association with PTKM (≈0.40), indicating that texts with richer local vocabularies also tend to exhibit higher modal particle density.
- **Pronouns.** PRON shows only weak correlations with all other variables, indicating that pronoun usage is largely independent of lexical diversity, text length, sentiment intensity, and MP usage.

Figure 1 visualizes all associations, highlighting weak, moderate, and strong correlations.

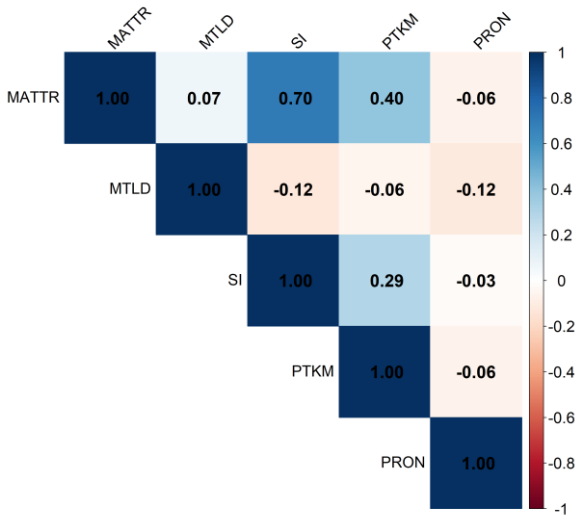


Figure 1: Correlation Heatmap

The correlation patterns motivate a series of multiple regression models that treat each feature in turn as a dependent variable. We report here the main patterns rather than every coefficient.

- **Predicting PTKM from MATTR, SI, and PRON.** The model is statistically significant but explains only about 10% of the variance ($R^2 \approx 0.10$). MATTR is the dominant positive predictor; SI is statistically significant but substantively negligible; PRON does not contribute meaningfully. This suggests that MP density is associated with lexical richness but may influenced by additional, unmodeled factors such as genre or artist style.
- **Predicting SI from PTKM, MATTR, and PRON.** This model has substantially greater explanatory power ($R^2 \approx 0.53$). Both MATTR and PTKM are positive and strong predictors, and the overall model accounts for over half of the observed variation in SI, which is notable for corpus-linguistic and social-science data. PRON is also a statistically significant positive predictor, but its effect size is small relative to MATTR and PTKM. These results indicate that songs with higher lexical diversity and higher MP density are more emotionally intense.
- **Predicting MATTR from PTKM, SI, and PRON.** The model explains roughly 54% of variance ($R^2 \approx 0.54$). PTKM is the strongest

positive predictor, pointing to a close coupling between MP use and lexical richness. SI is statistically significant but substantively small once scaling is considered. PRON shows a small negative association with MATTR: texts with higher first-/second-person prominence tend to be slightly less lexically diverse.

- **Predicting PRON from PTKM, SI, and MATTR.** This model has negligible explanatory power ($R^2 \approx 0.005$). Some predictors are statistically significant due to the large N, but their effect sizes are trivial. Pronoun variation appears driven by factors outside the present feature set (e.g. narrative perspective, genre conventions, artist-specific style).

Taken together, these results indicate a robust triangle of associations linking lexical diversity (MATTR), modal particles (PTKM), and sentiment intensity (SI): lexically richer texts tend to use more MPs and exhibit higher affective density, and MP-rich texts tend to be more emotionally intense. Pronouns play a marginal role in this structural-affective nexus.

Regarding lexical diversity measures, the divergence between MATTR and MTLD implies that not all diversity metrics are equally informative for lyrics with strong repetition and chorus structures. Our key RQ2 findings are therefore conditional on MATTR's behavior as a locally sensitive measure; MTLD appears less aligned with sentiment intensity in this genre, and alternative diversity metrics may yield partly different patterns. We treat this as a limitation and an avenue for future research rather than as evidence that one measure is universally superior.

3.2 Temporal trends (RQ3)

To address RQ3, we fit separate linear models with publication year as the predictor and each feature as the dependent variable. These models are intentionally simple; they are not intended as causal models but as first-pass summaries of diachronic trends.

1. **Lexical diversity (MATTR ~ YEAR).** We find a significant negative effect of year on MATTR ($\beta = -0.003857$, $p < 2e\text{-}16$, $R^2 \approx 0.21$). Lexical diversity decreases steadily over time, with modern songs using relatively simpler vocabularies. In relative terms, this corresponds to an approximate decrease of 1 – 2% per year, though precise percentage estimates depend on scaling and should be interpreted with caution. This pattern aligns with earlier findings for English lyrics (Parada-Cabaleiro et al., 2024), suggesting a cross-linguistic trend towards simplification.
2. **Modal particle density (PTKM ~ YEAR).** PTKM shows a small but highly significant negative slope ($\beta = -0.0001675$, $p < 2e{-}16$), with year explaining about 7 – 8% of the variance ($R^2 \approx 0.075$). MPs occur slightly less often in more recent songs, consistent with a shift towards more direct or pragmatically "lighter" language. Given the low mean and modest R^2, this trend should be interpreted as a weak but reliable tendency rather than a dramatic change.
3. **Sentiment intensity (SI ~ YEAR).** Sentiment intensity declines significantly over time ($\beta = -6.092e{+}10$, $p < 2e\text{-}16$, $R^2 = 0.14$). Relative-change estimates suggest a stronger proportional decrease than for MATTR or PTKM, though these percentages are sensitive to the absolute scale of SI. Substantively, the model indicates that songs have become somewhat less affectively dense, even if polarity remains relatively stable.
4. **Pronoun index (PRON ~ YEAR).** The model indicates a small increase in first-/second-person pronoun prominence over time ($\beta = 0.040649$, $p = 3.13e\text{-}7$, $R^2 = 0.0039$), but with very low explanatory power. While statistically robust, the effect is tiny and by itself insufficient to support strong claims about increased personalization. We therefore interpret it cautiously, as a weak trend that complements the structural changes observed in MATTR, PTKM, and SI.

Overall, the diachronic analyses suggest that, within the examined feature space, contemporary songs exhibit measurable temporal change: lexical diversity, modal particle density, and sentiment intensity all decline, while direct address (PRON) slightly increases (Figure 2). Together, these trends point towards a gradual simplification and mild "flattening" of emotional intensity, coupled with modest increases in subjectivity or listener orientation. However, given the modest R^2 values and the possibility of genre, artist, and topic effects, these interpretations should be seen as indicative rather than definitive.

A descriptive breakdown of personal pronouns in the corpus shows a clear dominance of first- and second-person forms. The most frequent item by a large margin is *ich* (255,260 instances), followed by *du* (126,909), while other pronouns such as *wir* (52,525), *es* (46,574), and *sie* (33,579) occur considerably less often. This distribution suggests that song lyrics in the corpus are strongly oriented toward self-reference and direct address, with a particular emphasis on the speaker's perspective. However, since these figures are aggregated across the entire time span, they do not by themselves indicate which pronouns drive the observed diachronic increase in the PRON index. A more fine-grained temporal analysis of individual pronouns is therefore left for future work.

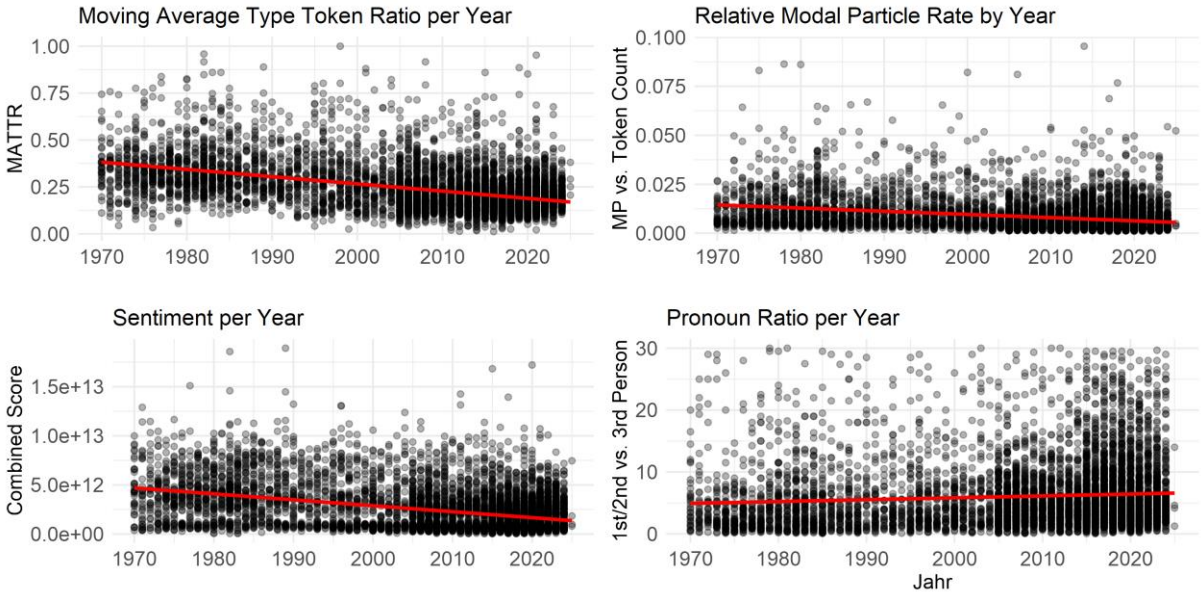


Figure 2: Linear Model Scatterplots of Temporal Trends in Sentiment and Linguistic Features

4. Conclusion and Outlook

This study has presented a multidimensional, empirically grounded analysis of linguistic and emotional patterns in Ger man pop lyrics across five decades, based on a large, longitudinally maintained and richly annotated corpus. Using Songkorpus, it has been sho wn how features such as lexical diversity, pronoun usage, modal particle density, and sentiment intensity can be operationalized and combined to explore how structural and affective dimensions co-vary over time. The resulting patterns point to potentially meaningful developments, including shifts in affective expression. However, the extent to which such patterns can be interpreted as evidence of broader cultural tendencies – such as forms of emotional flattening – remains necessarily limited. In particular, the interaction between linguistic form, genre conventions, and modeling assumptions makes it difficult to disentangle cultural change from representational and methodological effects. The analysis therefore provides a structured empirical perspective on these dynamics rather than definitive claims about their cultural significance.

Since the analytical payoff of this study lies in making core dimensions of lyrical communication measurable at corpus scale, the methods, models, and evaluation protocols are made openly available on the Songkorpus website[1], establishing a transparent foundation for replication and enabling extensions across languages, genres, and communicative contexts.

From a methodological perspective, the study demonstrates that the automatic identification of linguistically complex features such as modal particles and sentiment intensity can achieve a level of reliability sufficient for exploratory large-scale analysis when supported by targeted validation. The construction of a dedicated gold standard for modal particles, together with a focused sentiment annotation experiment, indicates that model performance approaches the range of human agreement observed for this type of data. At the same time, these findings underline the importance of cautious interpretation: even where annotation quality is comparatively high, the inferential step from measured linguistic features to broader claims about affect or cultural change remains non-trivial. Accordingly, the primary contribution of this study lies in demonstrating how such features can be modeled and evaluated in a challenging domain, thereby enabling more robust future investigations of creative, non-standard language.

Substantively, the analysis identifies robust associations between lexical richness, modal particle density, and sentiment intensity: lexically more diverse songs tend to use more modal particles and exhibit higher affective density. Diachronically, lexical diversity and modal particle use decline, while sentiment intensity also decreases and first- and second-person pronoun prominence increases slightly. These trends suggest a gradual shift towards simpler, somewhat less emotionally charged, and modestly more personalized language in German pop lyrics. They resonate with earlier findings for

[1] https://songkorpus.de/data/

English lyrics and invite further cross-linguistic comparison.

Several caveats and directions for future work follow from our findings, especially with regard to how feature design and validation can be integrated into corpus-processing pipelines to make large, evolving corpora more robustly analyzable over time:

- Our measure of "emotional dynamics" is based on global song-level sentiment intensity and does not capture within-song shifts, mixed polarities, or specific emotion categories. Segment-level modeling and multi-dimensional emotion classification would provide a richer picture of affective structure.
- Lexical diversity behaves differently across measures (MATTR vs. MTLD), especially in a genre characterized by strong repetition. Future work should systematically compare diversity metrics in lyrics and other poetic texts to clarify what aspects of variation they capture.
- The present models are correlational and largely linear; they do not establish causal direction. More advanced approaches – such as genre-stratified models, hierarchical or mixed-effects models, or causal modeling frameworks – could better disentangle the roles of genre, artist, and time.
- Our sentiment analysis relies on a lexicon-based baseline. Transformer-based sentiment and emotion models, fine-tuned on lyrics or related domains, could complement or challenge our results. Evaluating such models on large lyric corpora, ideally in combination with extended human annotation, is a promising avenue for future work.
- Finally, broadening the feature set to include syntactic complexity, rhythmic patterns, lexical fields, and discourse structure would further enhance the account of how linguistic form and affective content jointly shape the expressive texture of popular music.

Despite these desiderata and challenges, the present study demonstrates how large-scale, linguistically annotated corpora can reveal subtle relationships between language structure, affective expression, and cultural change. It provides a transparent methodological framework and open materials that are readily extensible to other languages and communicative contexts, while also illustrating how carefully validated automatic annotations enable empirically grounded analyses of culturally specific text types.

From the perspective of large-scale corpus management, the study directly engages with key challenges central to the CMLC agenda. The Songkorpus comprises texts that are "written to be sung" and therefore systematically diverge from the assumptions underlying most corpus-based modeling approaches, such as stable orthography, consistent segmentation, and standardized grammatical structure. This poses particular difficulties for diachronic analysis, as observed variation may reflect both linguistic change and evolving conventions of textual representation.

At the same time, the corpus highlights issues of access and sustainability: as a collection of copyrighted song lyrics, its availability is necessarily restricted, with implications for reproducibility and the transferability of analytical workflows. The approach adopted here addresses these constraints by focusing on derived representations (e.g., aggregated linguistic and emotional features) and by implementing preprocessing and modeling strategies that are robust to non-standard variation. In this way, the study demonstrates how creative, performance-oriented language can be systematically integrated into large-scale corpus analyses.

5. Bibliographical References

Abercrombie, G., Rieser, V., and Hovy, D. (2023). Natural Language Processing with Intra-Annotator Agreement. *ArXiv*, Preprint 2301.10684. https://arxiv.org/pdf/2301.10684

Bello, A., Ng, S.-C., and Leung, M.-F. (2023). A BERT Framework to Sentiment Analysis of Tweets. *Sensors*, 23(1), 506. https://doi.org/10.3390/s23010506

Bestgen, Y. (2024). Measuring Lexical Diversity in Texts: The Twofold Length Problem. *Language Learning*, 74: 638–671. https://doi.org/10.1111/lang.12630

Blühdorn, H. (2019). Modalpartikeln und Akzent im Deutschen. *Linguistische Berichte*, 259. Hamburg: Buske. 275–317. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-91746

Broll, S. and Schneider, R. (2023). Empirische Verortung konzeptioneller Nähe/Mündlichkeit inner- und außerhalb schriftsprachlicher Korpora. *Journal for Language Technology and Computational Linguistics*, 36(1). 113–150. https://doi.org/10.21248/jlcl.36.2023.240

Covington, M. and McFall, J. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17, 94–100. https://doi.org/10.1080/09296171003643098

Dahary, S., Edana, A., Apartsin, A., and Aperstein, Y. (2025). From Joy to Fear: A Benchmark of Emotion Estimation in Pop Song Lyrics. *ArXiv* Preprint 2509.05617. https://arxiv.org/pdf/2509.05617

Diewald, G. (2007): Abtönungspartikel. In Hoffmann, L. (Ed.), *Handbuch der deutschen Wortarten*. Berlin, New York: de Gruyter. 117–142. https://doi.org/10.1515/9783110217087.117

Emerson, G. and Declerck. T. (2014). SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, 30–38, Dublin, Ireland: Association for Computational Linguistics and Dublin City University. https://doi.org/10.3115/v1/W14-5805

Fehle, J., Schmidt, T., and Wolff, C. (2021). Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, Düsseldorf. https://doi.org/10.5283/EPUB.50833

Guhr, O., Schumann, A.-K., Bahrmann, F., and Böhme, H. J. (2020). Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In N. Calzolari et al. (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 1627–1632. Marseille: European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.202/

Hentschel, E. (2010). Partikelprofile literarischer Texte. In T. Harden, E. Hentschel (Eds.), *40 Jahre Partikelforschung*. Tübingen: Stauffenburg, 97–118.

Hunke, T., Huber, F., and Steffens, J. (2025). The Evolution of Song Lyrics: An NLP-Based Analysis of Popular Music in Germany from 1954 to 2022. *Music & Science*, 8. https://doi.org/10.1177/20592043251331155

Koch, P. and Oesterreicher, W. (2012). Language of immediacy – Language of distance: Orality and literacy from the perspective of language theory and linguistic history. In C. Lange, B. Weber, and G. Wolf (Eds.), *Communicative spaces: Variation, contact, and change*, 441–473. Frankfurt: Lang. https://doi.org/10.15496/publikation-20415

Kraus, M., Feuerriegel, S. (2019). Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. E*xpert Systems with Applications*, 118, 65-79, https://doi.org/10.1016/j.eswa.2018.10.002

Kreyer, R. and Mukherjee, J. (2007). The style of pop song lyrics: a corpus-linguistic pilot study. *Anglia*, 125(1). 31–58. https://doi.org/10.1515/ANGL.2007.31

Landis, J. R. and Koch, G. G. (1977): The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. https://doi.org/10.2307/2529310

Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge, University Press. https://doi.org/10.1017/CBO9781139084789

Malvern, D., Richards, B., Chipere, N., and Durán, P. (2024). *Lexical Diversity and Language Development. Quantification and Assessment*. London : Palgrave Macmillan. https://doi.org/10.1057/9780230511804

Mccarthy, P. and Jarvis, S. (2010). Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42, 381–92. https://doi.org/10.3758/BRM.42.2.381

Öhman, E. (2021). The Validity of Lexicon-based Sentiment Analysis in Interdisciplinary Research. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, 7–12, NIT Silchar, India. NLP Association of India (NLPAI). https://aclanthology.org/2021.nlp4dh-1.2/

Pang, B., Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Now Foundations and Trends. https://doi.org/10.1561/1500000011

Parada-Cabaleiro, E., Mayerl, M., Brandl, S., Skowron, M., Schedl, M., Lex, E., and Zangerle, E. (2024). Song lyrics have become simpler and more repetitive over the last five decades. *Scientifc Reports*, 14 (5531). https://doi.org/10.1038/s41598-024-55742-x

Schneider, R. (in preparation). Beyond Standard: Intelligent Modeling of Creative Language Using the German Song Corpus. In L. Herzberg, C. Mair, and A. Witt (Eds.), *Corpus linguistics 2040: Which data, which methods, which models?* Digital Linguistics, 6, Berlin/Boston: De Gruyter.

Schneider, R. (2026). Linguistic resources for the study of pop culture. In Valentin Werner, Cecilia Cutler, and Andrew Moody (Eds.), *Handbook of Language and Pop Culture*. Berlin, Boston: De Gruyter Mouton.

Schneider, R. (2020). A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with the Multi-Layer Annotated "Songkorpus". In N. Calzolari et al. (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 835–841. Marseille: European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.105/

Schoonjans, S. (2018). Modalpartikeln als multimodale Konstruktionen. Berlin/Boston: deGruyter. https://doi.org/10.1515/9783110566260

Scott, M. (2004). WordSmith Tools Version 4.0. Oxford: Oxford University Press.

Storø, S. R. (2022). Die Annotation der Modalpartikeln im GeWiss-Korpus. Eine syntaktische und semantisch-pragmatische Analyse der PTKMA-Annotation. *Deutsche Sprache. Zeitschriften für Theorie, Praxis und Dokumentation*, 22 (2), 124–149. https://doi.org/10.1515/ds-2022-0014

Sluyter-Gäthje, H. and Trilcke, P. (2022). Poesie als Fehler. Ein 'Tool Misuse'-Experiment zur Prozessierung von Lyrik. In *Proceedings 8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd)*, Potsdam. https://doi.org/10.5281/zenodo.6328201

Trippel, T. (2025). Metadata for research data. In P. Bański, U. Heid, and L. Herzberg (Eds.), *Harmonizing language data: Standards for linguistic resources*. Digital Linguistics, 4, Berlin/Boston: De Gruyter 251–279. https://doi.org/10.1515/9783112208212-011

Werner, V., Hiramoto, M., and Flanagan, P. (2025). Language and pop culture. Setting the agenda, *Journal of Language and Pop Culture*, 1(1). 1–17. https://doi.org/10.1075/jlpop.24034.wer

6. Language Resource References

Kudo, T. (2005). *CRF++: Yet Another CRF Toolkit.* http://taku910.github.io/crfpp/

Schneider, R. (2025). *Songkorpus - Linguistic Corpus of German Song Lyrics*. Release 7.0. https://songkorpus.de

The Infrastructure Behind Latvian National Corpora Collection

Roberts Dargis[1, 2], **Baiba Valkovska**[2]
University of Latvia, Raina bulvaris 19, Riga, Latvia
Institute of Mathematics and Computer Science, University of Latvia, Raina bulvaris 29, Riga, Latvia
{roberts.dargis, baiba.valkovska}@lumii.lv

Abstract

The rapid advancement of digital humanities and Natural Language Processing (NLP) necessitates centralized access to high-quality, large-scale language resources. This paper presents the technical infrastructure and evolving ecosystem of Korpuss.lv, the central access platform for the Latvian National Corpora Collection (LNCC). The LNCC consolidates 42 corpora developed by 14 institutions, comprising 2.8 billion tokens of written and spoken Latvian across diverse genres and annotation layers. Korpuss.lv has evolved from a simple metadata index into a comprehensive digital infrastructure that enhances corpus discoverability, accessibility, and usability for researchers in linguistics, digital humanities, and natural language processing. The platform integrates noSketchEngine as its primary corpus analysis tool and extends its functionality with custom modules, including a metadata-driven Corpora Explorer, a client-side Federated Content Search system, and precomputed UD-based Word Sketches. The ecosystem is further supported by CLARIN DSpace repositories for persistent storage and citation management, as well as a federated academic authentication architecture built on SATOSA and Keycloak via the CLARIN Service Provider Federation. The paper outlines architectural decisions, integration strategies, and future development plans.

Keywords: corpus infrastructure, federated authentication, corpus engine, CLARIN, Latvian corpora

1. Introduction

The rapid advancement of Natural Language Processing (NLP) and data-driven research in the digital humanities relies heavily on the availability of high-quality, large-scale language resources. For linguists, lexicographers, and AI developers, access to diverse and well-annotated corpora is essential for studying language variation, training models, and conducting quantitative textual analysis. However, language resources are often developed by independent institutions within separate, short-term projects. This frequently results in a fragmented digital ecosystem in which valuable data remain isolated, difficult to discover, or inaccessible due to incompatible formats and complex licensing restrictions.

To address these challenges, national and international infrastructures have emerged to consolidate language resources into centralized, standardized, and easily searchable platforms. For languages with smaller speaker populations and limited resources, such as Latvian, combining efforts within a unified national infrastructure is particularly important in order to maximize the impact and visibility of available data. A centralized ecosystem not only improves the discoverability of corpora but also provides standardized, user-friendly tools for querying and analyzing texts, reducing the need for extensive technical expertise among end users while properly managing academic access rights.

The Latvian National Corpora Collection (LNCC) is a diverse collection of corpora representing both written and spoken language. The collection contains 2.8 billion tokens of high-quality data totaling nearly 10 GB and covers a wide range of text types and genres, including news articles, social media posts, blogs, books, scientific texts, debates, and essays. The LNCC is a multi-institutional and multi-project initiative supported by the digital humanities and language technology communities in Latvia. Currently, it includes 42 corpora developed by 14 institutions.

A website called Korpuss.lv was developed to facilitate access to LNCC metadata and related services. Over time, Korpuss.lv has expanded in functionality, been extended with additional software components, and integrated with external resources. This paper focuses on the technical aspects of Korpuss.lv and its related ecosystem.

2. Related Work

Several projects are similar in scope. The most comparable are the CLARIN DSpace repositories (Straňák et al., 2020) established in various countries. CLARIN stands for Common Language Resources and Technology Infrastructure. A standard CLARIN DSpace installation hosts metadata records describing language resources and tools, and many of these records also provide access to the underlying data.

Some national CLARIN initiatives have expanded further or merged with related national activities, such as LINDAT/CLARIAH-CZ (Hajič et al., 2022), the Language Bank of Finland (The Language Bank of Finland, 1996), and Språkbanken CLARIN (Borin

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 44–48. Palma, Mallorca. May 11, 2026.

et al., 2012). These infrastructures have a broader scope, providing not only data but also computational infrastructure and tools for working with both hosted and user-supplied data.

The scope and focus of Korpuss.lv are to provide a curated collection of corpora together with browser-based tools preloaded with data, supporting the most common research use cases in linguistics and digital humanities. For additional use cases, links to download sites are provided for corpora that are available for download, allowing users to process the data with their own tools.

One of the most important components of the Korpuss.lv ecosystem is the corpus engine. Several open-source web-based corpus engines exist, each with its own strengths and limitations, such as noSketchEngine (Kilgarriff et al., 2014), Kontext (Machálek, 2020), and Korp (Borin et al., 2012). We use noSketchEngine because, when it was first deployed in 2017, it was the most feature-rich open-source web-based corpus engine available. Although alternative solutions have since matured, our user community is familiar with noSketchEngine, and migrating to a different platform is currently not a viable option. In the future, we may consider running an additional corpus engine in parallel if it offers distinctive features not available in noSketchEngine and demonstrates clear demand among users.

3. Korpuss.lv

Korpuss.lv[1] is a centralized digital infrastructure that hosts the LNCC. Initially developed as a simple index linking to external corpus platforms, it has evolved into an ecosystem comprising multiple interconnected components designed to improve corpus discoverability within Latvian research communities. The platform provides a user-friendly interface with filtering and sorting tools, federated content search across multiple corpora, and UD-based Latvian word sketches.

Our development strategy prioritizes the integration of established software solutions rather than building systems from scratch. However, as new use cases emerged for which no suitable ready-made tools were available, we developed additional modules to complement Korpuss.lv.

Korpuss.lv is implemented using Django[2], a high-level Python web framework. Python is widely used among NLP researchers, which facilitates the integration of existing NLP libraries when needed.

[1] https://korpuss.lv/en/

[2] Django framework – https://www.djangoproject.com/

3.1. Corpora Explorer

The Corpora Explorer serves as the main entry point to Korpuss.lv and presents the complete LNCC index through a filterable and sortable interface. Metadata-based filtering operates along three orthogonal dimensions: modality, distinguishing written text from speech; corpus type, distinguishing general-purpose from domain-specific corpora; and annotation level, including morphological, syntactic, error, manual, or diachronic annotation layers. Additionally, corpora with shared thematic provenance are grouped under labels such as historical, literary, or newspaper collections. This classification scheme is informed by conventions established by the Czech National Corpus project (Machálek, 2020) and the CLARIN resource family taxonomy (Fišer et al., 2018). The interface supports sorting chronologically by earliest data, reverse chronologically by most recent data, and by date of last update.

Each corpus is represented by a card displaying its identifier code, full name, and developing institution or institutions. The individual corpus page provides extended metadata, including associated publications, recommended citation formats derived from persistent identifiers assigned by the CLARIN DSpace repository (Section 5), and links to external download locations for corpora distributed under open or academic licenses.

3.2. Federated Context Search (FCS)

The Federated Content Search component enables simultaneous querying across all registered noSketchEngine corpus endpoints, returning both absolute and relative frequencies of the search term for each corpus. This functionality is particularly useful for identifying corpora that contain rare linguistic phenomena before performing detailed concordance analysis.

FCS is implemented as a client-side JavaScript application that sends requests directly from the user's browser to individual corpus engine endpoints, bypassing the Korpuss.lv application server. This architecture eliminates backend load and reduces latency associated with server-side proxying, providing a responsive asynchronous user experience as results arrive incrementally. The architectural trade-off is that all participating endpoints must expose permissive Cross-Origin Resource Sharing (CORS) headers. Endpoints without CORS support would require proxying through an intermediary server. All noSketchEngine instances within the Korpuss.lv ecosystem are configured accordingly. Result aggregation is performed in the browser. For each endpoint, the application issues a CQL query, parses the JSON response, and renders per-corpus frequency statistics in a unified table.

3.3. Word Sketches

The Word Sketch service[3] provides collocation and grammatical relation profiles for lemmas occurring in UD-parsed corpora. Unlike the dynamic word sketch computation available in the commercial Sketch Engine, the Korpuss.lv implementation relies on precomputed data structures. This design choice enables near-instantaneous query responses and minimizes runtime computational load.

Sketches are derived from dependency parses produced according to the UD annotation scheme (de Marneffe et al., 2021). The underlying data model is organized hierarchically across three levels: word (lemma), relation (typed syntactic dependency such as `nsubj`, `obj`, or `amod`), and collocation (a co-occurring lemma within a given relation). Sparsity and annotation noise are mitigated through a two-stage pruning strategy. Collocations with fewer than 10 occurrences are removed, and any relation node without remaining collocations, as well as any lemma node without remaining relations, is recursively removed. The resulting data are serialized and indexed for efficient lookup by lemma and part-of-speech tag.

4. Corpus Engine

The primary analytical tool within the Korpuss.lv ecosystem is the corpus engine. We use noSketchEngine, an open-source corpus management and analysis platform designed to support exploration of large text collections through a web interface. It represents a limited version of the commercial Sketch Engine system and is built on core components including Manatee for indexing and fast retrieval, Bonito for the graphical interface, and Corpus Query Language (CQL) for advanced search operations. The platform enables complex concordance searches, frequency list generation, and timeline-based analysis of language change. Users can perform simple searches by word form, lemma, or part-of-speech tag, or construct more advanced CQL queries combining multiple linguistic attributes and metadata filters.

Although it does not include automated word sketches or dictionary-building tools available in the commercial version, noSketchEngine provides robust support for investigating syntactic structures, semantic relationships, discourse patterns, and diachronic variation, making it well suited for research in linguistics, lexicography, and digital humanities.

We provide access to various types of corpora through our noSketchEngine instance, including learner corpora, parallel corpora, and speech event corpora. To ensure a consistent user experience, we aim to morphologically annotate all corpora that are not manually annotated using the same morphological annotator (Paikens et al., 2024). We have also automatically morphologically annotated a phonetic corpus containing word-level phonetic annotations (Auziņa et al., 2024). noSketchEngine uses tab-separated vertical files, a format also supported by the morphological annotator. For corpora containing manually annotated layers such as phonetic transcriptions, we convert the original data into vertical format using a custom processing pipeline and then generate morphological feature columns using the annotator. In rare cases, annotation accuracy may be slightly reduced when tokenization does not fully match the internal tokenization of the morphological annotator. However, this limitation is outweighed by the benefit of integrating morphological and manually annotated layers within the same corpus.

We are also exploring the use of Universal Dependencies (de Marneffe et al., 2021) and are gradually applying UD parsing to the corpora (Znotiņš, 2026). Although noSketchEngine does not natively support querying over dependency relations, the inclusion of UPOS and deprel layers provides valuable information for linguistic research, even for languages with relatively free word order such as Latvian.

noSketchEngine does not allow users to view entire documents when they exceed the maximum context window. This limitation makes it possible to provide broad access even when corpora are distributed under academic licenses with copyright restrictions. For corpora strictly restricted to academic users, we operate a separate noSketchEngine instance protected by academic authentication.

We plan to develop a dedicated document viewer that will be linked from corpus metadata within noSketchEngine. This viewer will expand research possibilities, especially for multimodal corpora containing images, audio, or video, such as speech, sign language, or OCR corpora. Academic authentication will also be supported in the document viewer for restricted corpora.

5. CLARIN

CLARIN (Common Language Resources and Technology Infrastructure)[4] is a European digital infrastructure providing sustainable access to language data and tools. It connects certified centers across multiple countries and offers repository services as well as federated authentication mechanisms. Within the Korpuss.lv ecosystem, CLARIN fulfills two essential technical roles: persistent and citable

[3] LVK2022 Word Sketches – https://korpuss.lv/skices

[4] CLARIN – https://www.clarin.eu

storage of language resources, and controlled access to restricted datasets through a federated identity framework.

The Service Provider Federation connects CLARIN-registered service providers to national identity federations across EU member states. This model allows academic users to access password-protected resources using their home institution credentials without registering separate accounts for each service. For providers, it ensures that access can be restricted to verified academic users while offering single sign-on convenience to end users through their existing institutional login.

CLARIN data repositories serve as the persistent storage and cataloging backbone of the infrastructure. They host structured metadata records describing corpora and tools, assign persistent identifiers to datasets, and often distribute data files directly. Persistent identifiers are essential for scholarly reproducibility because they enable unambiguous citation of specific corpus versions. When data are distributed under academic licenses, access is mediated by the federated authentication framework, which verifies institutional affiliation before granting download permissions. The most widely used repository platform within the CLARIN network is CLARIN DSpace[5].

Latvia operates its own national CLARIN DSpace instance (Skadiņa et al., 2020). The Korpuss.lv backend harvests metadata and associated persistent identifiers to automatically generate standardized citation recommendations on corpus detail pages. If data are hosted in the repository, a download button linking to the repository is also displayed.

6. Academical Authentication

Several corpora within the LNCC are restricted to academic users affiliated with recognized institutions. To enforce these restrictions, we required a federated authentication solution compatible with eduGAIN[6], the interfederation service linking national research and education identity federations. Direct membership in eduGAIN is not available to individual organizations, and participation must be mediated through a national federation. For this purpose, we use the CLARIN Service Provider Federation.

Typically, each web application requiring federated access must be registered independently as a service provider, which involves administrative overhead and manual review. To avoid registering each Korpuss.lv component separately, we implemented a single identity proxy registered once with the CLARIN Service Provider Federation that forwards authentication to downstream applications. For the application-facing layer, we use Keycloak[7], an open-source identity and access management platform responsible for session management and token issuance. However, Keycloak does not natively support SAML discovery services and cannot automatically ingest metadata from external identity provider registries, requiring manual configuration of each upstream provider.

To address this limitation, we introduced an intermediate proxy between Keycloak and eduGAIN that supports dynamic discovery and presents itself to Keycloak as a single unified identity provider. We evaluated SimpleSAMLphp[8] and SATOSA[9], selecting SATOSA primarily because it is implemented in Python, aligning with the broader Korpuss.lv technology stack and reducing operational complexity. In this architecture, SATOSA manages SAML discovery and federates outward to eduGAIN, while presenting a single SAML identity provider endpoint to Keycloak, which secures individual Korpuss.lv services. This layered design allows a single federation registration while maintaining flexibility to add new protected services with minimal configuration effort.

7. Conclusion

This paper presented the technical infrastructure and evolving ecosystem of Korpuss.lv, the central access point for the Latvian National Corpora Collection. By consolidating diverse corpora from multiple institutions into a unified platform, Korpuss.lv has significantly lowered the barrier to entry for researchers in linguistics, digital humanities, and natural language processing. Its evolution from a simple directory to a comprehensive digital ecosystem ensures that Latvian language resources are discoverable, accessible, and sustainable.

The development of Korpuss.lv demonstrates the value of integrating robust open-source solutions such as noSketchEngine and CLARIN DSpace with custom modules tailored to specific research needs. The implementation of an academic authentication gateway based on SATOSA and Keycloak effectively bridges institutional identity federations such as eduGAIN with secure access to academically licensed corpora.

The steady growth in user engagement and increasing academic citations referencing Korpuss.lv and the LNCC highlight the platform's importance within the research landscape. The next major development step is the creation of a dedicated document viewer with academic authentication support,

[5] CLARIN DSpace – https://github.com/ufal/clarin-dspace

[6] eduGAIN – https://edugain.org/

[7] Keycloak – https://www.keycloak.org/

[8] SimpleSAMLphp – https://simplesamlphp.org/

[9] SATOSA – https://github.com/IdentityPython/SATOSA

enabling full-text access and improved support for multimodal corpora containing audio, video, and images. Continued refinement of the Korpuss.lv ecosystem will further empower researchers and contribute to safeguarding the digital future of the Latvian language.

8. Acknowledgements

This work was supported by the EU Recovery and Resilience Facility project Language Technology Initiative (2.3.1.1.i.0/1/22/I/CFLA/002) in synergy with the State Research Programme Letonika – Fostering a Latvian and European Society project Digital Resources and AI Technologies for the Sustainability of the Latvian Language (DigiLATE) (VPP-IZM-LETONIKA-2025/1-0004).

9. Bibliographical References

Ilze Auziņa, Normunds Grūzītis, Roberts Darģis, Guna Rābante-Buša, Didzis Goško, Jānis Vempers, Raivis Kivkucāns, and Artūrs Znotiņš. 2024. Recent latvian speech corpora for linguistic research and technology development. *Baltic Journal of Modern Computing*, 12(4):646–658.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp — the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 474–478, Istanbul, Turkey. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Hajič, Eva Hajičová, Barbora Hladká, Jozef Mišutka, Ondřej Košarko, and Pavel Straňák. 2022. Lindat/clariah-cz: Where we are and where we go. *CLARIN. The infrastructure for language resources. Berlin: De Gruyter*.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1:7–36.

Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Pēteris Paikens, Lauma Pretkalniņa, and Laura Rituma. 2024. A computational model of Latvian morphology. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 221–232, Torino, Italia. ELRA and ICCL.

Inguna Skadiņa, Ilze Auziņa, Normunds Grūzītis, and Artūrs Znotiņš. 2020. Clarin in latvia: From the preparatory phase to the construction phase and operation. In *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN)*, pages 342–350.

Pavel Straňák, Ondřej Košarko, and Jozef Mišutka. 2020. Clarin-dspace repository at lindat/clarin. *Grey Journal (TGJ)*, 16.

The Language Bank of Finland. 1996. Kielipankki – The Language Bank of Finland. https://www.kielipankki.fi. University of Helsinki and CSC – IT Center for Science.

Artūrs Znotiņš. 2026. Pretraining and benchmarking modern encoders for Latvian. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*. ACL. To appear.

Optimized for AI: Curating the Icelandic Gigaword Corpus for Stable LLM Training

Jón Friðrik Daðason, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
{jon.fridrik.dadason, steinthor.steingrimsson}@arnastofnun.is

Abstract

The Icelandic Gigaword Corpus (IGC) is a primary resource for Icelandic NLP, with its current version containing 2.7 billion words of curated text. The IGC is traditionally distributed in a TEI-XML format, a hierarchical structure that allows for rich linguistic annotation and metadata. However, this format introduces significant friction for modern machine learning workflows. Even high-quality curated corpora have been found to contain "unwanted" text sequences—such as fragmented lists or repetitive boilerplate that may trigger instabilities during training of large language models. In this paper, we present a new processing pipeline designed to optimize the IGC for AI development. We describe a filtering approach focusing on training stability, including fuzzy deduplication to reduce the risk of data leakage, with the aim to provide high-quality data for stable model convergence. Furthermore, we introduce a new JSONL distribution format that bridges the gap between TEI-XML and machine-actionable data, facilitating easier access and safer training for models aiming to work with Icelandic.

Keywords: LLM training data, Large corpora, Filtering

1. Introduction

Pre-training corpora for modern language models increasingly rely on web-crawled data. Although abundant, such data often contains substantial amounts of low-quality text and duplicate content. A manual evaluation of five multilingual web-crawled corpora found that 15 out of 205 language-specific subsets did not contain a single usable sentence, and that the proportion of usable text was below 50% in 87 subsets (Kreutzer et al., 2022). As a result, web-crawled corpora are typically filtered and deduplicated before pre-training. At the same time, Transformer-based language models have proven remarkably robust to noisy pre-training data. Filtering and deduplication often yield only modest average gains on downstream tasks, and the benefit of deduplication in particular is inconsistent across corpora and model scales (Raffel et al., 2020; Muennighoff et al., 2023).

The Icelandic Gigaword Corpus (IGC; (Steingrímsson et al., 2018; Barkarson et al., 2022)) is a curated, high-quality monolingual corpus for Icelandic, currently containing approximately 2.7 billion running words across a wide range of genres. In recent years, it has become the primary resource for pre-training and fine-tuning language models for Icelandic (Snæbjarnarson et al., 2022; Daðason and Loftsson, 2022; Daðason, 2025). Given the modest downstream impact often reported for text quality filtering and deduplication, it is reasonable to ask whether such preprocessing is necessary for a curated corpus such as the IGC. However, downstream performance is not the only consideration.

Recent work shows that even a small number of low-quality examples within a single training batch (e.g., highly repetitive n-gram sequences) can trigger immediate pre-training instability, potentially causing the model to plateau at worse performance or diverge during training (Walsh et al., 2025). Our own experience pre-training on the IGC suggests a similar pattern, where examples consisting primarily of non-running text (e.g., tabular content or lists of names, dates, or monetary amounts) or text in a foreign language appear to disproportionately contribute to training instability, despite representing only a small fraction of the corpus.

In this paper, we present a filtering and deduplication pipeline for the IGC, and we present a new JSON Lines (JSONL) distribution format for the corpus, optimized for LLM pre-training and fine-tuning, to be published alongside the TEI format.

2. Related Work

Text quality filtering is often performed using rule-based methods in which numerical features are extracted from text and compared against predetermined thresholds. Text is discarded if any feature falls outside an acceptable range. A major benefit of rule-based methods is that they are highly explainable, making them well-suited to settings where minimizing false positives is a priority. That said, there are no standard rulesets for text quality filtering, and the choice of rules and thresholds varies considerably between works.

The web-crawled C4 corpus (Raffel et al., 2020) used a comparatively simple heuristic filtering pipeline with both line-level and document-level rules. Lines were removed if they were very short,

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 49–56. Palma, Mallorca. May 11, 2026.

lacked a terminal punctuation mark, or matched boilerplate patterns (e.g., “terms of use”, “privacy policy”, or “cookie policy”). Documents were discarded if they contained certain strings indicating quality issues (e.g., “lorem ipsum”, “Javascript”, or “{”), or if a language classifier did not label them as English with high confidence.

The MassiveWeb corpus (Rae et al., 2022) applied document-level filtering based on a range of features. Documents were discarded if their word count or mean word length fell outside an acceptable range, or if they contained a high proportion of lines beginning with a bullet point, a high ratio of hash symbols or ellipses to words, a low proportion of words containing at least one alphabetic character, too few unique stop words, or a high proportion of repeated lines, paragraphs, or n-grams.

FineWeb (Penedo et al., 2024) reused many of the rules applied in C4 and MassiveWeb. In addition, documents were discarded if they had a high ratio of short lines, a high proportion of characters in duplicated lines, or a high proportion of lines ending without a terminal punctuation mark.

Despite the diversity of features and large rulesets applied in prior work, Daðason and Loftsson (2024) found that relatively few features had the greatest impact on text quality classification. Evaluating 13 commonly used features on a manually labeled dataset, they found that an unsupervised classifier achieved its highest F_1 score using only three: perplexity, stop word ratio, and mean subword length. The filtering pipeline described in this paper takes a similarly targeted approach, with rules derived directly from inspection of low-quality content found in the IGC rather than from prior work on web-crawled data.

The Text Encoding Initiative (TEI) guidelines (TEI Consortium, 2026) remain the standard for the long-term preservation and linguistic annotation of national corpora. They allow for nested structures and granular metadata with detailed information on everything from part-of-speech tags to licensing, provenance, and how the text was sourced and cleaned. TEI is used by national corpora such as the British National Corpus (Burnage and Dunlop, 1992), the Bulgarian National Corpus (Koeva et al., 2010), and the National Corpus of Polish (Przepiórkowski et al., 2008).

Although rich in metadata, the hierarchical complexity of TEI-XML introduces significant overhead for machine learning workflows. Consequently, recent projects have prioritized machine-actionable formats such as JSONL to facilitate high-throughput, stream-based data loading. The Norwegian Colossal Corpus (Kummervold et al., 2022) and the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021) are notable examples of this trend. Furthermore, other projects employ slightly different approaches to obtain the same goals. For example, the ParlaMint project (Erjavec et al., 2023), containing 17 parliamentary corpora in 16 main languages, is published both in the TEI-based ParlaMint format and a TSV-format that can easily be converted to JSONL or another format suitable for machine learning (ML) or NLP. While TEI will remain the master format for the IGC, we introduce a JSONL distribution alongside it to support language model pre-training and fine-tuning.

3. The IGC

Since first being released in 2018, the Icelandic Gigaword Corpus (IGC, Steingrímsson et al., 2018) has undergone significant expansion, with updated versions published annually or biannually. The corpus has grown from an initial 1.2 billion running words to its current iteration of approximately 2.7 billion words (Barkarson et al., 2022; Barkarson and Steingrímsson, 2024). The IGC is a tagged and lemmatized corpus. While news media constitutes the largest share of the data, the corpus is diverse, as seen in Table 1

The corpus is distributed under a dual-licensing scheme that balances open-access goals with copyright restrictions from various content providers. Approximately 63% of the corpus is available under a permissive license, CC BY 4.0, allowing for unrestricted use and redistribution. The largest categories under this license are *social media and internet forums* (∼30%), specific news outlets (∼17%) and parliamentary records (∼10%). The remaining 37% falls under the IGC Custom License, a more restrictive license which permits research use and training of language models but prohibits republication of raw texts. All downloads are centralized through the Árni Magnússon Institute for Icelandic Studies.

The IGC’s adoption has been broad, spanning corpus linguistics as well as modern NLP. On the linguistic side, it has been used to track language change, frequency distributions, and usage patterns, including as a resource underlying the maintenance of the Database of Modern Icelandic Inflection (Bjarnadóttir, 2012; Bjarnadóttir et al., 2019). In NLP, it has, e.g. served as a source for back-translation to generate synthetic parallel data for machine translation (Símonarson et al., 2021; Jasonarson and Steingrimsson, 2025), and as the core pretraining corpus for Icelandic encoder models (Snæbjarnarson et al., 2022; Daðason and Loftsson, 2022). Most recently, it provided training, validation, and test data for a language modeling task at GKÍ2026, the Icelandic AI competition[1].

[1]See: https://github.com/gervikeppnin/GKI2026/tree/main/golden_plate_on_thingvellir_NLP

Category	Running Words	(%)
Adjudications	79,625,568	2.95%
Law, bills and resolutions	60,623,312	2.24%
Published books	13,824,783	0.51%
Scientific/academic journals	20,894,101	0.77%
News media	1,442,810,126	53.43%
Parliamentary proceedings	266,115,169	9.85%
Social media	806,949,613	29.88%
Wikipedia	9,718,240	0.36%
Total	**2,700,560,912**	**100.00%**

Table 1: Distribution of running words and percentages across different text categories in the 2024 version of the IGC.

Despite this wide adoption, the IGC's primary distribution format presents a practical obstacle for LLM training. The corpus is encoded in TEI (Text Encoding Initiative) XML (TEI Consortium, 2026), a standard in corpus linguistics that provides rich structural and linguistic annotation. For language model training, however, this richness becomes friction: the deep nesting of XML tags means that raw text must be extracted through custom preprocessing pipelines, typically converting the corpus to plain text or JSONL while carefully preserving document boundaries and discarding markup. An unannotated TEI version introduced in 2022 reduces some of this overhead by providing cleaner sentence-level nodes, but the fundamental conversion step remains unavoidable, and doing it correctly is non-trivial. This paper addresses that gap directly, presenting a version of the IGC that has been processed and formatted specifically for LLM training.

4. The Processing Pipeline

In this section, we describe the processing pipeline used to filter and deduplicate the IGC. As the corpus contains a very low proportion of truly low-quality documents, we opted for a rule-based approach to text quality filtering in order to minimize false positives (i.e., erroneously removing high-quality text). Documents were first normalized, and boilerplate text was removed where possible. Documents containing issues that were not easily corrected, or that indicated deeper quality problems, were discarded entirely. Finally, although empirical evidence for the benefits of deduplication on downstream performance remains inconclusive, we applied fuzzy deduplication to reduce the risk of data leakage between training and validation splits.

4.1. Boilerplate Text

Many documents in the IGC contain strings that are irrelevant to their main content. This includes navigational elements, keywords, categories, lists of related articles, advertisements, social media links, and metadata. Additionally, we removed certain references to elements not present in the plain-text version of the documents, such as embedded videos or audio. We also removed frequently used signatures and author bylines, as their presence might bias models towards generating the same signatures or names in their output. We constructed a separate ruleset for each subcorpus, with each rule consisting of a regular expression pattern or a specific substring to be removed.

4.2. Escaped Elements

Some documents in the IGC contain escaped HTML or XML elements, such as > and & (representing the characters > and &, respectively). We unescaped all documents containing such elements. This sometimes required multiple unescape operations (e.g., first converting a doubly-escaped ampersand from & to &, and then to &). Some documents used custom escaped elements, which we unescaped by either inferring the appropriate form from the name of the element or by reviewing correctly rendered versions of the affected documents.

4.3. Character Normalization

Some documents contain nonprintable or otherwise undesirable characters, primarily private use area Unicode characters (code points reserved for private use) and certain ASCII control characters. We removed or normalized these characters as appropriate. Additionally, due to their rarity, we normalized all Unicode whitespace characters (e.g., no-break space and thin space) to either a literal space or a newline character.

4.4. Whitespace Normalization

Documents in the IGC have generally been preprocessed by collapsing multiple adjacent whitespace characters into one, stripping leading and trailing spaces from each line, and removing empty

lines. However, a small number of documents do not fully conform to this, and earlier steps in our pipeline may also have introduced or eliminated whitespace or left certain lines empty. We therefore repeated this whitespace normalization step to ensure a consistent input format.

4.5. Short Documents

It is common practice to remove short documents from pre-training corpora, although there is no standard approach to doing so (Raffel et al., 2020; Rae et al., 2022; Ettinger et al., 2025). Following Rae et al. (2022) and Penedo et al. (2024), we removed documents containing fewer than 50 words.

4.6. Stop Word Ratio

We calculated the ratio of stop words to all alphanumeric tokens within each document. Documents with a very low stop word ratio tend to consist primarily of non-running text (e.g., tabular data, lists, and bullet points), or foreign-language or non-linguistic content. We discarded documents whose stop word ratio fell below a minimum threshold of 22%.

4.7. Internal Duplication

We discarded documents in which at least 20% of sentences are duplicated. This can occur in short news articles that open with a brief preview, which is then repeated verbatim in the body of the article. For short articles, this preview can account for a significant portion of the total document length. Beyond this pattern, some documents contain unexpected instances of duplicate text, either present in the source file from which the text was extracted or mistakenly introduced during the extraction process itself. As it is difficult to determine which duplicate segments to remove, if any, we chose to discard documents with a high degree of internal duplication.

4.8. Code

Some documents contain unintended code, such as HTML, JavaScript, or XML, typically introduced when text is extracted from malformed source files. These are generally short and incomplete snippets that can be difficult to remove cleanly and may indicate deeper quality issues within the document. For this reason, we discarded any document containing code. Documents that intentionally contained code, such as Wikipedia articles on programming languages or computer science topics, were excluded from this filter.

4.9. Character Encoding Errors

Character encoding errors can be introduced when documents are decoded using an incorrect character encoding. For example, decoding a UTF-8-encoded file as Latin 1 will produce garbled output, such as rendering “ö” as “Ã¶”. Certain characters are strongly associated with such errors, and we discarded any document in which they were present.

4.10. Phrase Filtering

We identified and removed several classes of unwanted documents using targeted string matching. These include messages requesting that the reader log in, create an account, or subscribe in order to view the full contents. Documents containing such strings are often cut off mid-sentence, featuring only a heavily truncated version of the full text. We also removed documents containing warnings that JavaScript must be enabled to view the content, as these may indicate deeper quality issues. Finally, we discarded documents containing phrases closely associated with template-generated content, such as Wikipedia disambiguation pages.

4.11. Optical Character Recognition Errors

Some subcorpora in the IGC contain documents digitized using optical character recognition (OCR) software. On rare occasions, this process yields heavily garbled results, with a significant proportion of non-alphanumeric symbols that rarely occur in high-quality text. As correcting such errors is difficult, time-consuming, and risks introducing additional errors, we discarded documents where such symbols are present.

4.12. Old Documents

Documents published prior to 1930 often contain non-standard spelling and were thus discarded.

4.13. Fuzzy Deduplication

We performed fuzzy deduplication using MinHash and LSH with 20 bands of size 13. For each cluster of near-duplicate documents, we retained the longest document from the subcorpus whose documents had the highest overall pass rate on the preceding filtering steps, using this as a proxy for subcorpus quality.

5. The JSONL Format and Publication

To facilitate high-throughput training, we provide the processed IGC in JSONL format, as shown in

```
{
  "tei_archive": "IGC-News2-22.10.TEI.zip",
  "tei_path": "IGC-News2-22.10.TEI/fotbolti/2005/02/IGC-News2-fotbolti_1260328.xml",
  "source": "fotbolti.is",
  "altered": true,
  "text": "..."
}
```


Figure 1: Example of the new JSONL distribution format for the IGC. The inclusion of `tei_path` ensures full traceability to the original TEI-XML source.

Figure 1. Each line in the JSONL distribution represents a document in the corpus. It contains the raw text alongside essential provenance metadata, including whether it has been altered by the processing pipeline, as indicated in the "altered" field. This allows researchers to track any document back to its original TEI-XML source if the rich linguistic annotations (such as POS tags or lemmas) are required.

The IGC's dual-licensing model necessitates a split distribution strategy:

Open Access (CC BY 4.0): The portion of the corpus under permissive licensing is made available both in the CLARIN-IS repository[2] and on Hugging Face[3], which allows for immediate integration into standard ML data loaders. This part of the corpus is ≈795M running words after filtering.

Custom License: The other part of the corpus, published with the custom license which allows for research and training but restricts raw text redistribution, is only available through the CLARIN-IS repository[4]. This part of the corpus is ≈895M running words after filtering.

Users wishing to train on the full running 1.6 billion words of filtered data can easily merge the two JSONL streams. To support reproducible research, we also provide pre-defined training and validation splits for both portions of the corpus.

6. Discussion

Table 2 shows the number of documents and tokens discarded during the filtering and deduplication process. These statistics do not include documents from published books, journals, or social media, which are distributed as shuffled paragraphs or sentences for licensing and copyright reasons, making them unsuitable for inclusion in a pre-training corpus. The remaining text consists of 1.9 billion tokens, of which 1.6 billion are words, across 4.8 million documents. The most commonly discarded category by document count consists of documents containing fewer than 50 words, although these account for a negligible proportion of tokens in the IGC. The largest category by number of tokens is near-duplicates, totaling approximately 84 million tokens. The most common type of near-duplicates were news articles published both online and in printed newspapers. The stop word ratio filter flagged a substantial number of documents with a high proportion of non-running text. Although such documents represent a small proportion of the corpus, we believe their removal should result in measurably improved training stability. Overall, while a substantial number of documents were removed, most were simply unsuitable for pre-training rather than truly low-quality or noisy text.

Table 3 shows the number of documents that were normalized or altered, excluding those discarded entirely. Aside from boilerplate removal, these categories proved to be quite rare. Nevertheless, since we prioritized high accuracy with minimal false positives, we expect that even these infrequent corrections will help prevent the model from wasting capacity on malformed or noisy text.

These results strongly suggest that text quality in the IGC is generally very high, as expected for a curated corpus, but also confirm that a number of documents contain pathological text sequences that can negatively impact training stability. In practice, we observed that training examples consisting primarily of foreign-language text or non-running text sometimes caused spikes in gradient norms during pre-training, leading to increased training and validation losses, as described by Walsh et al. (2025). Although there is a subjective element to decisions about what text should be normalized or discarded, monitoring training stability can provide empirical grounding for such choices.

That said, filtering certain types of text in favor of training stability can come at a cost. For a monolingual model, it would be unrealistic to expect any benefits from cross-lingual transfer by retaining the small amount of foreign-language text present in

[2] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/381

[3] https://huggingface.co/datasets/arnastofnun/IGC-2024-filtered-1

[4] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/382

Category	Documents	Tokens
Short documents	348,841	12,777,006
Near-duplicates	215,553	83,956,512
Stop word ratio	175,923	38,791,290
Internal duplication	41,737	27,445,929
Code	18,382	7,552,647
Character encoding errors	7,047	1,321,170
Phrase filtering	7,039	1,142,784
OCR errors	6,799	3,745,532
Old documents	2,240	17,772,286
Total (unique)	**752,784**	**183,021,842**

Table 2: Number of documents and tokens for each category discarded in the filtering pipeline. Near-duplicates do not include the canonical document in each cluster.

Category	Documents
Boilerplate text	429,108
Escaped elements	10,996
Character normalization	3,212
Whitespace normalization	480
Total (unique)	**443,298**

Table 3: Number of documents normalized or otherwise altered by category, not including discarded documents.

the IGC. However, filtering out non-running text risks discarding meaningful information. For example, tables convey a great deal of structured content, but have historically been difficult to represent in plain text format. An alternative would be to encode such content in a format better suited for subword tokenizers and language models, such as JSON or Markdown. However, this would need to be performed at the text extraction stage, prior to filtering, and we therefore leave it to future work. Until then, discarding such text in favor of improved training stability remains the most practical option.

While a great deal of foreign-language text can be filtered using simple heuristics like stop word ratios, language classifiers offer a much more robust approach in principle. Documents that are primarily in other languages are excluded from the IGC, but foreign text often appears within Icelandic documents, sometimes as lengthy excerpts. For future versions of the pipeline, we plan to experiment with paragraph-level language identification. Fedorova et al. (2026) compare GlotLID (Kargaran et al., 2023) and two versions of OpenLID (Burchell et al., 2023), finding F_1 scores in the high 0.90s on hard evaluation sets and very close to 1 for other evaluation sets when disambiguating between Scandinavian languages. While they do not evaluate Icelandic, Burchell et al. (2023) report an F_1 score of 1.0 for Icelandic in their results for the first version of OpenLID. Incorporating a language identifier based on one of these recent tools could therefore serve as a valuable addition to the pipeline.

7. Conclusions and Future Work

We have presented a new JSONL distribution of the IGC, in a format specifically optimized for LLM training. We also described our filtering and deduplication pipeline and demonstrated that even high-quality corpora can contain segments that are flawed and unwanted for most NLP tasks. This included repetitive templates, code snippets, and tabular data or other non-running text that can trigger instabilities when training LLMs.

To safeguard against such issues we used targeted, mostly rule-based filtering. By removing approximately 752,000 problematic or redundant documents, we provide a more robust foundation for stable training without sacrificing the richness and linguistic diversity that makes the IGC unique among Icelandic text corpora. Furthermore, by distributing the corpus in a machine-actionable JSONL format with clear training and validation splits, we lower the barrier to entry for researchers and developers working on Icelandic AI.

Future iterations of this work will focus on two primary directions. First, rather than discarding non-running text such as tables or lists, we aim to develop heuristics to reformat these structures where possible into more useful representations (e.g., Markdown or structured JSON). This would preserve information that is currently lost during text extraction. Second, we plan to integrate state-of-the-art language identifiers, such as OpenLID, at the paragraph level to better handle code-switching and foreign-language excerpts within Icelandic documents. Through these continued refinements, we aim to ensure the IGC remains the definitive resource for the next generation of Icelandic language technology.

Acknowledgements

This work was funded by the Icelandic Strategic Research and Development Programme for Language Technology 2025, grant no. 250251-5301.

8. Bibliographical References

Starkaður Barkarson, Steinþór Steingímsson, and Hildur Hafsteinsdóttir. 2022. Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *LREC 2012 Proceedings: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 – AfLaT*, pages 67–72.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland. Linköping University Electronic Press.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An Open Dataset and Model for Language Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Gavin Burnage and Dominic Dunlop. 1992. Encoding the British National Corpus. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: Design, Analysis and Exploitation*, volume 10 of *Language and Computers*, pages 79–95. Rodopi, Amsterdam.

Jón Friðrik Daðason. 2025. *Language Representation Models for Low- and Medium-Resource Languages*. Ph.D. thesis, Reykjavik University.

Jón Friðrik Daðason and Hrafn Loftsson. 2022. Pretraining and Evaluating Transformer-based Language Models for Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7386–7391, Marseille, France. European Language Resources Association.

Jón Friðrik Daðason and Hrafn Loftsson. 2024. Unsupervised Outlier Detection for Language-Independent Text Quality Filtering. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 383–393, Torino, Italia. ELRA and ICCL.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michal Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, et al. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. OpenLID-v3: Improving the Precision of Closely Related Language Identification – An Experience Report.

Atli Jasonarson and Steinthor Steingrimsson. 2025. AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama Models for Low Resource Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 695–704, Suzhou, China. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language Identification for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Svetla Koeva, Diana Blagoeva, and Siya Kolkovska. 2010. Bulgarian National Corpus Project. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani,

Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling Data-Constrained Language Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*.

Adam Przepiórkowski, Rafał L. Górski, Barbara Lewandowska-Tomaszyk, and Marek Łaziński. 2008. Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 Submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword Corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

TEI Consortium. 2026. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 4.11.0 edition. Oxford, Providence, Charlottesville, Nancy. Last updated on 18th February 2026. Accessed February 2026.

Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, et al. 2025. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.

9. Language Resource References

Starkaður Barkarson and Steinþór Steingrímsson. 2024. Icelandic gigaword corpus (IGC-2024ext) - unannotated version. CLARIN-IS.

Starkaður Barkarson, Steinþór Steingrímsson, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Finnur Ágúst Ingimundarson, and Árni Davíð Magnússon. 2022. Icelandic gigaword corpus (IGC-2022) - unannotated version. CLARIN-IS.

Hellenic National Corpus: the current state

Maria Gavriilidou[1], **Nikos Sidiropoulos**[1]

Institute for Language and Speech Processing, Athena Research Center
Athens, Greece
(maria, nsidir)@athenarc.gr

Abstract

The Hellenic National Corpus (HNC) is an integrated online environment offering access to standard Modern Greek language material and to related analysis tools. The HNC corpus has been developed in two main phases, and currently comprises over 97 million words exclusively of written language, sourced from printed resources or scraped from the internet. The material has been automatically lemmatized and morphologically annotated, while a subset of 100,000 words has been further manually corrected, in order to produce a freely downloadable error-free corpus. Through the dedicated platform, the users have access to concordances, morphological analysis of words and statistical information (frequency) at word, lemma, part of speech and n-gram levels. Future steps include the expansion of the material in both historical and coverage dimensions: the inclusion of material from older phases of the language is foreseen, as well as the addition of dialectal material besides standards language.

Keywords: Greek corpus, access environment, analysis tools

1. Introduction

The Hellenic National Corpus (HNC)[1] is an integrated online environment offering access to Modern Greek language material and to related analysis tools. The HNC corpus currently comprises over 97 million words exclusively of written language, sourced from printed resources or scraped from the Internet. The corpus material has been automatically lemmatized and morphologically annotated. Through the dedicated platform, the users have access to KWIC-type or full sentence concordances, statistical information (frequency) and collocations at word, lemma and part of speech levels. Recent improvements of the HNC concern both the language material and the platform: quantitative and qualitative enrichment of the language material; addition of new text types and genres from the internet (digital press, social media, blogs, etc.); development of a new backend (document uploading platform and metadata editor); improvement of existing and addition of new search functionalities; redesign of the HNC user interface; improved visualizations of results; and, finally, creation of the Golden Part of Speech Tagged corpus, an automatically annotated and manually corrected small corpus subset, freely available through the CLARIN:EL National Infrastructure for Language Resources and Technologies[2] (Gavriilidou et al. 2024).

The current paper is structured as follows: Section 2 presents an overview of the historical evolution of the HNC; Section 3 describes the criteria and methodology for corpus collection, classification and processing across phases; Section 4 elaborates on the HNC Platform and its current functionalities; Section 5 focuses on the Golden Part of Speech Tagged corpus; and finally Section 6 concludes with future steps.

2. Historical Evolution of HNC

HNC was developed in two broad phases: the first phase (starting in 1992) set the objectives of the endeavor, specified the collection criteria and the methodology, and proceeded to collect the material and develop the platform for user access (Hatzigeorgiu et al., 2000, Gavrilidou, 2002). This phase resulted in the first version of HNC, which contained 42 million words. The largest portion of the material originated from Newspapers (Table 1), which were exclusively printed at that time, while the internet was critically under-represented. Publishing houses contributed literary and scientific works, which represented almost 10% of the material. These proportions, which were due to the availability (or scarcity) of the respective sources, rendered the corpus unbalanced.

Publication medium	
Book	9,4%
Internet	0,3%
Newspaper	61,3%
Periodicals	5,9%
Other	23,1%
Total	100,0%

Table 1: First phase proportions by Publication medium

Every single text that formed part of the HNC was accompanied by the appropriate license agreement. The license agreements which were signed with the publishers provided the material with strict restrictions, namely: HNC was allowed to include excepts but not the entire text provided

[1] https://hnc.ilsp.gr/

[2] https://www.clarin.gr/en

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 57–62. Palma, Mallorca. May 11, 2026. ©ELRA Language Resources Association, 2026

by the sources, to offer to the users very restricted amount of text results, i.e., the sentence containing the user query term, plus the previous and the next sentence; provision of the whole paragraph or even more of the whole text was forbidden due to Intellectual Property Rights restrictions.

The HNC Platform hosting the corpus and providing access to the language material was also designed and developed during the first phase. Its functionalities included filtering the material to select specific texts according to criteria (e.g., only newspapers, texts of a specific author, publisher or date, etc.), in order to construct sub-corpora on which the search was performed. Content search focused on word, lemma and/or part of speech, and up to 3 combinations thereof; additionally, HNC offered word and lemma frequencies. In order to ensure lawful use of the material granted by the publishing houses, the platform was designed to allow only online access but no downloading of the material.

After a long period of maintenance of the HNC and user support but no addition of new material, the second phase (2020-2021) undertook the quantitative and qualitative enrichment of the corpus and the improvement of platform functionalities. The necessity for the enrichment of the corpus material was due firstly, to its small size for a national corpus, secondly, to the need to keep up with the technical evolution of the field, and thirdly, to cater for the new language production modes and language use established through the internet: in order to reflect digital-era language use, new genres and text types needed to be included, mainly born digital material (instead of digitized), social media material, etc. The text collection criteria and methodology were revised. A target size of 100 million tokens was set and new material was collected, processed, annotated and added to the HNC. The platform was redesigned as regards both backend and frontend, and it was enriched with additional search and analysis functionalities. These steps are detailed in the following sections.

3. Corpus Collection and Processing

3.1 Selection Criteria

HNC aimed to be a representative corpus of the current Greek language; therefore, the terminus post quem adopted was 1976, the year of the establishment of the standard Modern Greek language as the official language of education and public administration, which put an end to decades of diglossia. In order to focus on the current use of the language, most of the texts included have been produced from 1990 onwards, while a special exemption was made in the case of literature, where older significant and influential literary texts have also been included.

The issue of balance and representativeness has long concerned corpus linguists, in combination with the conflict between adherence to strict predefined design principles and the actual availability of sources. In the case of HNC, a practical approach was adopted rather than strict ratios between genders: the objective was to cover as many aspects of current language use as possible, through the inclusion of a large variety of genres, text types and topics. Given the focus on the standard language, dialectal material has not been included (geographical dialects as well as sociolects), as diverging from the standard. Readability was also used as text selection criterion: texts with high readability (high circulation newspapers, best-selling books) were preferred due to their influence in the evolution of the language.

In the initial phase, the collection strategy mainly consisted of requests to publishing houses and news agencies, frequently striving to overcome their reservations and skepticism, and convince them to sign provision agreements governing the lawful inclusion of their material in the HNC and the subsequent access provision for research purposes.

During the second phase, as mentioned in Section 2, expansion focused on including online-native content. For this, topic-focused web crawling techniques were used. Seed websites were selected after assessing their contribution towards both quantity and balance of content, based on the original selection criteria (standard, current, non-dialectal language, with a variety of genres and topics). The seed websites were fed to the ILSP Focused Crawler (Papavassiliou et al. 2013), a tool developed to automatically locate monolingual and bilingual texts of specific topics on the web. Initial identification of candidate texts for inclusion was carried out by this tool. The automatically gathered material was further screened for license: texts needed to be openly available for research purposes, without usage restrictions. The appropriate license was at least CC BY-NC-SA 4.0 (Creative Commons Attribution –NonCommercial–ShareAlike 4.0), which allows sharing for research purposes and adaptation with proper attribution. The detection of such licenses was first done automatically by the tool (via website disclaimers) and then manually reviewed. Entries linking to third-party content without appropriate licenses were excluded.

Duplicate entries (about 16%, mainly due to news agencies reposting each other) were automatically identified and only one version was included in the HNC.

Given that the second phase ended in 2021, the issue of synthetic data generated by LLMs (currently a hot issue for corpus creation), at the time had not yet appeared.

3.2 Text Classification

Every text included in the HNC corpus is accompanied by metadata providing bibliographic information (title, author, publisher, publication date), Publication Medium, Genre, and Topic. The typologies adopted in the first phase for Medium, Genre and Topic adhered to the specifications of EU project PAROLE (PAROLE, 1995), based on which many national corpora were documented in the years 1990-2000. One of these corpora was HNC, which, according to these specifications was an adequately sized corpus, especially for an under-resourced language as Greek at that time.

HNC was maintained throughout the following years, although no enrichments were possible. The second phase of the enrichment of HNC (2020-2021) aimed to benefit from technological advancements (greater computational capacity and storage), and to respond to wider social evolutions (increasing production of digital content, augmented use of social media, etc.). Consequently, it was considered necessary to focus on digital content sourced from the internet, which was not satisfactorily represented in HNC until then. Criteria for topic selection also needed updating, to reflect the digital linguistic production.

During the second phase of quantitative and qualitative enrichment, more than 65,000 new texts with almost 50 million words were added, sourced exclusively from the internet. This was dictated by the enrichment principles but was also enabled by the ease of access of digital textual material and clear licensing schemes. After the enrichment phase, digitally born material constitutes more than half of the material of HNC (55.74%), while digitized, originating as printed material totals 44.26%.

The new texts were selected to reflect a great variety of Genres (Table 2) and Topics (Table 3).

Genre proportions	
Opinion	26,2%
Information	62,8%
Official	1,3%
Scientific/Educational	2,1%
Private	0,2%
Literature	1,4%
Instruction	0,9%
Proceedings	0,1%
Discussion	5,0%
Miscellaneous	0,1%
TOTAL	100,0%

Table 2: Current Proportions by Genre

Topic proportions	
Society	49,6%
Economy	12,4%
Leisure	9,3%
Arts	8,7%
International issues	5,7%
Politics	5,6%
Health	4,3%
Sciences	3,7%
Culture	0,6%
Miscellaneous	0,3%
TOTAL	100,0%

Table 3: Current proportions by Topic

Genre and Topic classification in the second phase was performed semi-automatically: the initial classification was performed by the ILSP Focused Crawler based on relevant information identified in the text, followed by manual correction.

3.3 Text Processing

Before being added to the HNC, all texts went through three main stages: normalization, annotation (structural and linguistic), and metadata addition. Normalization removed non-textual elements and converted files into standard XML format. Structural annotation (segmentation and tokenization) identified structural elements such as paragraphs, sentences, and tokens (words, abbreviations, numbers, dates). Linguistic annotation included lemmatization and morphological annotation (part of speech tagging and morphological analysis). These processing stages were conducted using the ILSP Feature-based multi-tiered POS Tagger[3] (available through the CLARIN:EL infrastructure for Language Resources and Technologies); the tool's accuracy is 96.28% (Papageorgiou et al., 2000). Manual correction of the morphological annotation results has been performed exclusively for the Golden Part of Speech Tagged Corpus (see Section 5); manual correction of approximately 100 million words was considered not feasible. It must be noted that the process of normalization and annotation was applied also on the already existing material of the initial phase, to secure conformity.

4. The HNC Platform

The corpus is accessible to the users through a dedicated platform, allowing users to search for texts via filters based on the metadata described.

[3] http://hdl.handle.net/11500/ATHENA-0000-0000-23E8-3

Thus, users can select specific texts or define sub-corpora based on date of publication, author, topic or any of the metadata. Texts are accessed through the specially designed user interface, but are not available for downloading, due to the agreements signed with the copyright holders (publishing houses, institutions, etc.), that allow access to but not free distribution of their material.

The HNC platform consists of two web applications: the frontend interface which provides user access, and the backend interface which provides platform administration. Both applications were developed in PHP scripting language, and they are hosted in a Internet Information Server (IIS).

4.1 Technical Description: the Backend

The core of the HNC platform is the SQL Server Database, where data are stored. Queries to the database are performed via SQL Query language, the response is handled by PHP and presented through the frontend.

The backend interface caters for the preparation of the documents to be inserted to HNC. A web-based management environment was implemented in PHP and connected to the corpus database. Mass document insertion was achieved via a PHP script which loaded the contents of the XML files into the SQL Server Database that hosts the HNC.

The backend platform also included a metadata editor for single XML file editing, classification, and insertion. Through this editor, annotators inspect, validate or correct (if needed) the automatically added metadata identified by the crawler during the collection process; they can also add the appropriate metadata in missing cases. Each document's metadata can be edited through the respective form (Figure 1), while word annotation editing is also available (Figure 2).

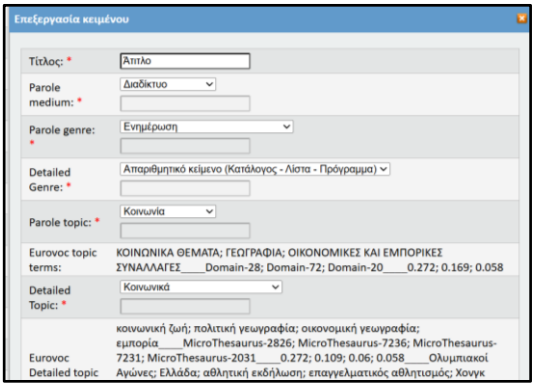


Figure 1: Document metadata editing

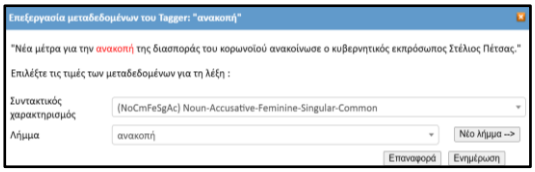


Figure 2: Word annotation editing

4.2 The Frontend: Search Functionalities

The frontend interface was designed to provide user-friendly interaction and easy customization in query building. Although the HNC frontend is open to use without registration, its main functionalities are available only to registered users. Guest users have limited results to their queries (50 sentences) while registered users have a significantly higher sentence limit (5,000 sentences). Furthermore, registered users can create and save sub-corpora to use in their searches; they also have access to the *Analysis* and *Correlation* functionalities (see Section 4.2), whereas these features are not available to non-registered users. This approach safeguards the performance of the HNC by protecting it from simultaneous multiple queries by large numbers of non-registered users.

The frontend interface allows users to perform filter-based selection of texts, to formulate queries to perform content search, to request morphological analysis of a word, or statistics of single or multiple items. Specifically, content search is achieved through word, lemma, or part of speech search, and combinations thereof (e.g. lemma X followed by lemma Y, lemma X followed by a preposition and then by a noun, etc.). The system retrieves sentences which comply with the search criteria and provides KWIC-type concordances (Figure 3) or full sentences (Figure 4) containing the requested query terms.

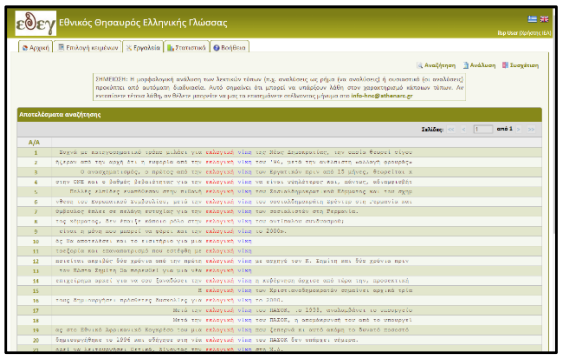


Figure 3: HNC concordance of a two-item query, each one marked in different color

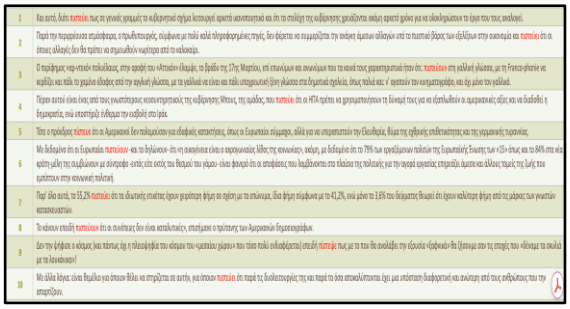


Figure 4: Full-sentence concordance

The users can also obtain statistical information, i.e., frequency of words, lemmas or parts of speech, n-gram frequencies, and lists with the most frequent words and lemmas in HNC.

The *Analysis* functionality offers the 'linguistic profile' of an item (word or lemma), namely, morphological analysis, frequent n-grams it participates in, and the yearly distribution of the item in HNC, i.e., its frequency on the time scale, based on the date of publication of the texts that contain the specific item (Figure 5).

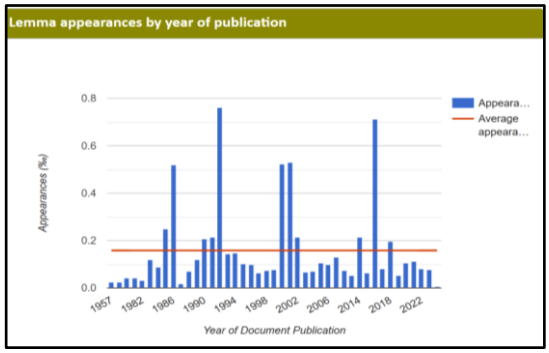


Figure 5: Yearly distribution of a lemma

Collocations a word participates in can be studied in detail through the most frequent words preceding or following a specific word (Figure 6).

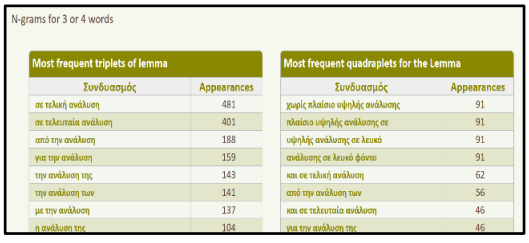


Figure 6: Most frequent n-grams for a lemma

The relation between two words is analyzed through the *Correlation* functionality, which provides information on their joint appearance in HNC: if and how frequently they appear together and how far apart (Figure 7) and their comparative frequency through the years (Figure 8).

Correlation & distance	
Joined appearances	498
Sequential appearances (% of joined ones)	0 (0 %)
Appearances in distance between 2 and 5 words (% of joined ones)	232 (46.59 %)
Appearances in distance between 6 and 10 λέξεις (% of joined ones)	72 (14.46 %)
Appearances in distances over 10 words (% of joined ones)	194 (38.96 %)
Minimum distance in sentences	2
Maximum distance in sentences	84
Average distance in sentences	11

Figure 7: Correlation & distance of two words

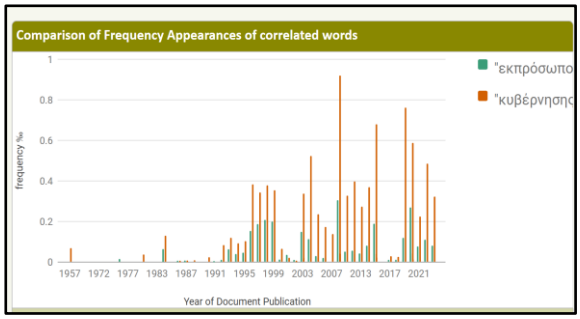


Figure 8: Joint yearly appearance of two words

Currently, registered users can store the sub-corpora they create. Query results are not stored, but search history is preserved along with its parameters so it can be replicated. Statistics, and specifically the top-frequency words/lemmas are also provided to registered users in CSV format for download.

5. The Golden Part of Speech Tagged Corpus

The Golden Part of Speech Tagged Corpus is a small subset of HNC (100,000 words), which consists of texts selected from a variety of sources covering various domains. Given that the material was planned from its conception to be freely distributed, the texts were selected based on their license (either CC0 4.0 or CC BY 4.0). The texts have been crawled from the web and underwent cleaning and removal of boilerplate material, manual correction of typos and spelling mistakes, automatic lemmatization and part-of-speech tagging for each word using the ILSP Feature-based multi-tiered POS Tagger, and manual correction of the results by linguists, in order to provide error-free material. The Golden Part of Speech Tagged Corpus is freely downloadable via CLARIN:EL as a single XML file[4].

[4] http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C

6. Future Steps

Although maintenance of HNC, as well as user support are taken care of, further enrichments or improvements have not been possible after the second phase, i.e. since 2021, due to the lack of funding.

With the proviso of funding availability, future steps concern the enrichment of the content with older material (diachronic expansion) and the addition of dialectal material (geographical and social dialects). Diachronic expansion will proceed stepwise, from the most recent to the older versions of the language, as this has also repercussions on the accompanying processing tools which need to be updated to successfully deal with older nominal and verbal inflection systems. As regards annotation tools, experimentation with LLMs is planned, focusing on the use of existing models for the processing of the existing material, with the aim to test their performance and to provide a new lemmatized and annotated version if appropriate. Additional steps include the finetuning of the existing models for the processing of the dialectal material which will be part of the corpus.

Issues to be tackled concern the identification and treatment of synthetic data and other machine-generated data such as translationese.

Finally, freely available material of the HNC is foreseen to be made available through the CLARIN:EL infrastructure, whose platform allows the downloading of both material and processing results.

7. Bibliographical References

Gavriilidou, M., Piperidis, S., Galanis, D., Pouli, K., Labropoulou, P., Bakagianni, J., Tsiouli, I., Deligiannis, M., Kolovou, A., Gkoumas, D., Voukoutis, L., and Gkirtzou, K. 2024. The CLARIN:EL infrastructure: Platform, Portal, K-Centre. In Lindén, K., Kontino, T. and Niemi, J (eds.) *Selected papers from the CLARIN Annual Conference 2023,* Linköping Electronic Conference Proceedings 210. ISBN: 978-91-8075-740-9. DOI: https://doi.org/10.3384/ecp210005

Gavrilidou, M. 2002. The Hellenic National Corpus on-line. In*: Revue belge de philologie et d'histoire.* Tome 80 fasc. 3, 2002. Langues et littératures modernes - Moderne taal en litterkunde. pp. 1003-1015. https://www.persee.fr/doc/rbph_0035-0818_2002_num_80_3_4652

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H., and Demiros, I. 2000. Design and implementation of the online ILSP Greek Corpus. In *Proceedings of Language Resources and Evaluation Conference* (LREC-2000). Athens, Greece. European Language Resources Association (ELRA) https://aclanthology.org/L00-1250/

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. 2000. A Unified POS Tagging Architecture and its Application to Greek. In *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC'00), Athens, Greece. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/181.pdf

Papavassiliou, V., Prokopidis, P., and Thurmair, P. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria. Association for Computational Linguistics, https://aclanthology.org/W13-2506/, pp. 43-51.

PAROLE. Design and composition of reusable harmonized written language reference corpora for European languages. 1995. Technical report, PAROLE Consortium. MLAP: 63-386

8. Language Resource References

Institute for Language and Speech Processing - Athena Research Center (2021). Golden Part of Speech Tagged Corpus. Version 1. [Dataset (Text corpus)]. CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C

Institute for Language and Speech Processing - Athena Research Center (2015). ILSP Feature-based multi-tiered POS Tagger. Version 1. [Software (Tool/Service)]. CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-23E8-3

Corpas Náisiúnta na Gaeilge 2022-2029: A Project Overview

Ó Meachair, M. J., Bhreathnach, Ú., Ó Raghallaigh, B., Ó Cleircín, G., Měchura, M., Scannell, K.

Dublin City University (DCU), Fiontar & Scoil na Gaeilge, Droim Conrach, D09 N920
{micheal.omeachair, una.bhreathnach, brian.oraghallaigh, gearoid.ocleircin, michal.boleslav.mechura}@dcu.ie, kscanne@cadhan.com.

Abstract

This paper reports the latest developments, planned works, and issues of the Corpas Náisiúnta na Gaeilge (henceforth: CNG, translation: *the National Corpus of Irish*) project, detailing the work that has been completed to date, current work, and planned future work. This report details the compilation of corpora, development of a project website and part-speech tagger, the challenges of expanding existing corpora, and the addition of historical and legal corpora. We also present the training and outreach activities of the project.

Keywords: corpus linguistics, NLP, low-resource language

The CNG project is being administered by the Gaois (www.gaois.ie) research group with funding from the Department of Rural and Community Development and the Gaeltacht and the National Lottery. Gaois is a research group in Fiontar & Scoil na Gaeilge, DCU, comprising lecturers, researchers and postgraduate students. Our aim is to sustain and transform Irish language and culture through the development of innovative and trusted resources. These resources include the National Terminology Database for Irish (www.tearma.ie), the Placenames Database of Ireland (logainm.ie), among others, and since 2024 this also includes the CNG project (www.corpas.ie). CNG built on the Corpus of Irish for Lexicography which was a proof-of-concept and is detailed in Ó Meachair, et al (2021).

1. Project Phases

In this section we introduce the three phases of the CNG project. This section concludes by reporting on a selection of challenges that arose with the compilation and provision of data.

1.1. Phase 1: 2022-2024

Corpus name	Size
Corpas Náisiúnta na Gaeilge, CNG (The National Corpus of Irish)	101 million words
Corpas Monatóireachta na Gaeilge, CMG (The Monitor Corpus of Irish)	1 million words *per annum*
Corpas na Gaeilge Labhartha, CGL (The Corpus of Spoken Irish)	9 million words
Corpas na Gaeilge Scríofa, CGS (The Corpus of Written Irish)	131 million words

Table 1: Corpora compiled during Phase 1

In brief, CNG is a balanced representative corpus of Irish language for the period 2000-2024. We included written data from both online and printed sources (for example: literature, news, academic, blogs), from as many genres and registers as possible. We also included spoken data from radio and television, from speeches and lectures, as well as creative spoken works such as songs and stage dramas.

CMG includes samples from genres that reliably publish in Irish every year: news, novels, legal texts, annual governmental reports and business reports.

CGL includes a variety of data from radio, television, and in-person contexts. Some of the data are transcribed and some are written to be spoken, such as scripts and lectures.

CGS includes written data that has gone through an editorial process, thus excluding blogs and social media posts, as well as some self-published documents. No balancing has been done to reduce the volume of legal texts or the more prolific news agencies.

During Phase 1 the www.corpas.ie website was developed, leveraging NoSketchEngine (Natural Language Processing Centre, 2025; Rychlý, P. 2007) for concordancing purposes, and a part-of-speech tagger was developed that built on UD-Pipe technologies (Straka and Straková, 2017) and used the PAROLE tagset for Irish (Uí Dhonnchadha, 2009).

1.2 Phase 2: 2025

Phase 2 lasted one year (2025) and saw delivery of the following outputs:

Outputs
Addition of 1-million words to CMG, for the year 2025.
Addition of 1-million words to CGL from the period 2000-2025

Generation of updated N-gram and word-frequency lists
Additions, improvements, and corrections to the tagging hitherto completed on all corpora

Table 2: Outputs completed during Phase 2

These outputs were added to the project website in Q1, 2026.

1.3 Phase 3: 2026-2029

Phase 3 commenced in January 2026, and has the following outputs as its stated objectives.

Output	**Metric**
Addition of the Parallel English-Irish corpus of legislation from Ireland and the European Union to the project website.	69.3m in Irish, 63.7m in English
Addition of a further four years, 4 million words, to CMG.	4-million words
Addition of collected data to CGS	Not stated
Addition of collected data to CGL	Not stated
Addition of the *Bailiúchán na Scol* (1937-1939) Collection to the project website.	10-million words of Irish
Improvements to search functionality and website features, and updates to word lists.	Subject to testing

Table 3: Planned outputs from Phase 3

1.4 Compilation and provision challenges

Our corpora depend on licences for the scientific usage of copyrighted data, some of which are long-standing while other new sources are consistently sought out. News agencies and publishing houses are the biggest contributors of data, and it is recognised that sharing their works is not a core part of their workflows. This is not an issue for low-resourced languages alone, as documented in Kupietz and Lungen (2014). This can relate to API access changing in some manner or to fiction and non-fiction publications being produced as stand-alone projects, rather than archived files complete with metadata. Furthermore, not all authors have the same contracts with their publishers and looking into these contracts is very time consuming for publishing houses.

We work hard to explain our scientific uses in plain non-technical language so as not to complicate this process, assuring other parties we cannot and will not re-share any publication in a way that would infringe on Irish copyright law. We do share some derived data via our project website (www.corpas.ie) or by special request - this is done in accordance with legal research in this area and both Irish and EU copyright law (Iacino, G. et. al. (2024), Kamocki, P. (2021), Directive (EU) 2019/790). Word-embeddings and n-gram data can be found on the project website. *Bone fide* researchers can contact the Gaois group to request a shuffled subset of the corpus data. This does not include spoken data (because it can be more personal and recognisable) creative works (novels, poetry, songs, *etc*); a consideration that reflects the respect we have for our creative colleagues and contributors in the dawning age of AI-generated publications. We recognise how the shuffling diminishes the context of the language use, but the legal and ethical considerations supersede this and we remind users context can be linguistically explored via www.corpas.ie. The shuffled text includes roughly one sentence per line and the shuffling is done automatically at least 1000 times (publication forthcoming).

2. Training and outreach

The authors of this paper have also contributed to training and outreach in a number of areas. A selection of which are introduced below.

2.1 Training

The module "Taighde Corpais" (Corpus Research) worth 10 ECTS and taught at NFQ Level 9 has been administered since 2022. This is a mandatory module for students completing an MA in Irish with Fiontar & Scoil na Gaeilge.

Seminars and social-media campaigns have also been run with a view to training the Irish-speaking community to both use the www.corpas.ie website and to engage in corpus-linguistic research.

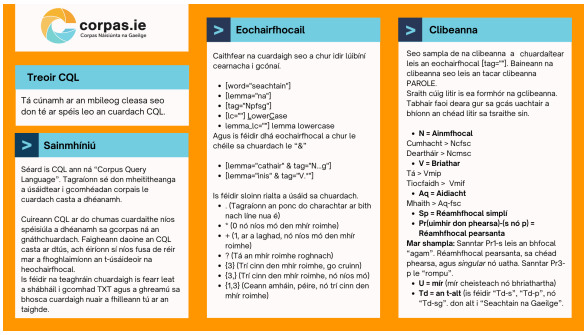


Figure 1. A cheat sheet for using CQL and PAROLE on www.corpas.ie.

2.2 Outreach and conferences hosted

The authors of this paper have been engaged in a number of outreach activities, inter-institutional publications, consultancy and provision of data for Irish-language NLP research, as well as participation in the CLARIN-ERIC network as

ambassador for corpus-linguistic research in Ireland.

2.2.1 Project launch[1] (November 29, 2024)

In November 2024 we hosted a series where guest speakers came to Fiontar & Scoil na Gaeilge, DCU to launch our project. Contextualisation of corpus use, use cases for national corpora in other countries, and technological developments were among the topics presented.

Figure 3. Michal Křen (Charles University, Prague) presenting the Czech National Corpus and its variety of uses

2.2.2 Corpus Linguistics for the Languages of Ireland (Nov 13-14, 2025)

The Gaois research group hosted workshops and a conference for researchers and practitioners working on corpus research for the languages of Ireland.

Workshops titles[2]:

- 'Common Language Resesources and Technology Infrastructure' - Dr. Martin Wynne (Oxford University)
- '*Quo Vadis* Corpas.ie' - Dr. Mícheál J. Ó Meachair & Dr. Michal Měchura.

Conference keynote speaker and contributors[3]:

- 'Corpus Linguistics and Language in Ireland: A promising future?' - Prof. Raymond Hickey.
- 19 researchers presented completed work and ongoing research.

3. Conclusions

We are working to stay true to the principles of the Gaois research group while delivering the CNG project and serving the Irish-language community. With a view to expanding our user base we will add to our training and outreach efforts by conducting and disseminating in-depth linguistic research, by adding specialized corpora and collections that are widely known, and by continuing to host conferences and workshops.

4. Bibliographical References

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

Iacino, G., Kamocki, P., Du, K., Schöch, C., Witt, A., Genêt, P. and Calvo Tello, J. (2024). Legal status of Derived Text Formats–2nd deliverable of Text+ AG Legal and Ethical Issues –. RuZ - Recht und Zugang. 5. 149-172. 10.5771/2699-1284-2024-3-149.

Uí Dhonnchadha, E. (2009) Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar. PhD thesis, Dublin City University..

Kupietz, M. and Lüngen, H. 2014. Recent developments in DeReKo. In Proceedings of the ninth conference on international language resources and evaluation (LREC'14), pages 2378–2385, Reykjavik, Iceland. ELRA.

Kamocki, P. (2021). When Size Matters. Legal Perspective(s) on N-grams. CLARIN Annual Conference. 122-128. 10.3384/ecp18014.

Natural Language Processing Centre (NLP Centre) at the Faculty of Informatics, Masaryk University (2025) "NoSketch Engine" *Available at:* https://nlp.fi.muni.cz/trac/noske. (Accessed 8 March 2026)

Ó Meachair, M. J., Ó Raghallaigh, B., Bhreathnach, Úna, Ó Cleircín, G., and Scannell, K. (2021). 'Tiomsú Corpais don Taighde Foclóireachta: Corpas Foclóireachta na Gaeilge (CFG2020)'. *TEANGA, the Journal of the Irish Association for Applied Linguistics*, 28, 278-305.

Rychlý, P. (2007) Manatee/Bonito-A Modular Corpus Manager. In: RASLAN. p. 65-70.

Straka, M., and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of CoNLL 2017.

[1] https://www.gaois.ie/en/blog/seoladh-corpas-ie

[2] Day one blogpost: https://www.gaois.ie/en/blog/an-teangeolaiocht-chorpais-la-1

[3] Day two blogpost: https://www.gaois.ie/en/blog/an-teangeolaiocht-chorpais-la-2

General Regionally Annotated Corpus of Ukrainian: Recent Developments and Future Plans

Maria Shvedova
National Technical University
"Kharkiv Polytechnic Institute"
Kyrpychova str. 2, 61002, Kharkiv, Ukraine
Friedrich Schiller University Jena
Fürstengraben 1 07743, Jena, Germany
mariia.shvedova@khpi.edu.ua

Abstract

The General Regionally Annotated Corpus of Ukrainian (GRAC) effectively serves as a national corpus. GRAC v.19 (2025) contains 2 billion tokens from over 800,000 texts (1816–2025). The corpus has multi-level annotations: rich metadata including regional tags, morphological annotation based on the VESUM dictionary, and partial semantic annotation. GRAC is the source of several derivative projects, including UD_Ukrainian_ParlaMint, ParaRook parallel corpora, Rada_Trees, and others.

Keywords: Ukrainian language; national corpus; regional annotation; morphological annotation; semantic annotation; corpus development; digital linguistic infrastructure

1. Introduction

The General Regionally Annotated Corpus of Ukrainian (GRAC) (Maria Shvedova (2017–)) is currently the largest manually curated and annotated corpus of Ukrainian.

When GRAC was initiated in 2016, Ukrainian corpus linguistics needed a larger, more comprehensively annotated resource to support varied research tasks and serve as national infrastructure. The latest version (GRAC.v.19, 2025) contains 2 billion tokens from over 800,000 texts by approximately 35,000 authors, covering modern Ukrainian from 1816 to 2025, including texts from both Ukraine and the diaspora. GRAC includes a wide range of text types: fiction (FIC), non-fiction (NOF), academic writing (ACA), journalism (JOU), official documents (OFF), ego-documents such as memoirs, diaries, and correspondence (EGO), transcribed public (SPU) and private speech (SPR), poetry (POE), and internet communication (ICM), drawn from printed, recorded, handwritten, and web sources (see Appendix A for proportions).

GRAC operates as an open volunteer project hosted at Jena University (Germany), with partial funding through university research grants (2019-2024), and contributed by students and specialists from multiple Ukrainian universities, including Kharkiv Polytechnic Institute, Ukrainian Catholic University, Kyiv-Mohyla Academy, and others. Since 2018, over 1,600 Ukrainian students from more than 15 universities have contributed to development through university courses and volunteer initiatives. Since 2024, GRAC has been receiving selected web texts from PAWUK, courtesy of IPI PAN (Kieraś et al., 2025).

Beyond serving as a reference corpus, GRAC functions as foundational infrastructure for specialized resources. Recent derivative projects include: a Universal Dependencies treebank *UD_Ukrainian_ParlaMint* (Shvedova and Lukashevskyi (2024-2025)) (Shvedova et al., 2025); *ParaRook* parallel corpora (Shvedova and Lukashevskyi (2023–)); *PluG* (copyright-free Ukrainian texts from GRAC, for download) (Shvedova and Lukashevskyi (2024)); *Rada_Trees* (parliamentary transcripts, 1990-2024, annotated with UDPipe2 and TagText) (Arsenii Lukashevskyi (2025)); *ParlMix-UA-RU* (parliamentary code-mixing dataset) (Olha Kanishcheva (2025)) (Kanishcheva et al., 2026); *ParaFarm* (English-Ukrainian multiple-translation corpus of George Orwell's *Animal Farm*) (Maslij (Kalashnyk) and Shvedova (2025)); and *PressMint-UA* (Ukrainian component within comparable corpora of historical newspapers, work in progress) (CLARIN ERIC, 2025).

This paper presents the annotation architecture underlying GRAC, discusses the methodological challenges encountered in its development, and outlines planned enhancements.

2. Pipeline and Technical Infrastructure

The general pipeline for updating GRAC comprises the following stages: text collection and metatextual annotation, text preprocessing, lemmatization, morphological tagging and semantic annotation, and corpus compilation.

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 66–70. Palma, Mallorca. May 11, 2026.

2.1. Text collection and metatextual annotation

Texts are collected through multiple resources: digitized printed sources (OCR with manual correction), transcribed audio recordings, and downloads from online sources including news portals (such as hromadske.ua, zaxid.net, procherk.info) and digital libraries (ukrlib.com.ua for fiction, libraria.ua for historical press, scc.knu.ua for doctoral theses, and others).

Metadata is entered manually into spreadsheets and subsequently transferred to a dedicated metadata database (developed by Sergey Yarygin), which enforces validation by checking that all values belong to predefined sets of allowed values and verifying the consistency between metadata records and text files. Text files are stored separately and are linked to their metadata records by filename.

All texts receive metatextual annotation including author information, dates of creation and publication, stylistic register, genre classification, information about the source medium and language of the original text.[1] A distinctive feature is GRAC's regional annotation, which tracks geographical variation through publication location and the author's region of origin. This is particularly important given Ukrainian's complex dialectal landscape and different historical varieties (Shvedova and von Waldenfels, 2021).

2.2. Text Preprocessing

Text preprocessing includes cleaning with the CleanText program[2] (Starko et al., 2021), which addresses a range of issues common in texts obtained via OCR or downloaded from the internet. These include erroneous apostrophe characters (replaced with the standard U+0027), Latin characters mixed into Cyrillic text (e.g., the Latin *i* substituting the Ukrainian *і*), digits used in place of visually similar letters, soft hyphens within words, and dangling or end-of-line hyphens that split words across lines.

Non-Ukrainian text (predominantly Russian) is excised and replaced with three hyphens (---). For parliamentary transcripts, this process has been automated using CleanText's language detection algorithm, which compares word counts matched against Ukrainian and Russian dictionaries. For other texts, the removal of non-Ukrainian fragments has in some cases been performed manually.

[1] https://uacorpus.org/en/rozmitka-tekstiv

[2] https://github.com/brown-uk/nlp_uk

2.3. Lemmatization, morphological tagging, and semantic annotation

Morphological annotation is performed automatically using the TagText program[2] based on the VESUM open-access dictionary (Starko and Rysin, 2022). Each token receives a lemma and a composite tag consisting of morphological features (part of speech, case, number, gender, tense, person, etc.) separated by colons. The format is *word/lemma/tag*, for example *korpusiv/korpus/noun:inanim:p:v_rod* (*korpusiv* 'corpus.GEN.PL', lemma *korpus*, noun, inanimate, plural, genitive). The system handles detailed morphological annotation using an extensive tagset and includes specialized tags for colloquialisms, archaisms, vulgarisms, and orthographic variants. Evaluation of the TagText tagger shows high precision: 99.3% for lemmas, 98.7% for pos, and 94.5% for full morphological tags. Specialized rule-based tools process non-standard orthography in historical texts, including the normalization of the Western Ukrainian Zhelekhivka orthographic system (Shvedova et al., 2021, 2022; Chemerys et al., 2023).

Morphological ambiguity presents an ongoing challenge. Most GRAC versions retain all possible analyses for ambiguous forms, though GRAC.v.17a and GRAC.v.19a implement automatic disambiguation. Following Rysin's description of the system (Shvedova et al., 2025), the disambiguation in TagText operates on three levels. First, rarely used word forms are discarded: for example, *rozpalenij* ('FIRE-PST.PASS.PTCP-ADJ-F.LOC.SG') could theoretically be parsed as an imperative verb form (*rozpalenij* 'INFLAME-IMP.2SG'), but is almost always an adjective and is tagged as such. Second, rule-based disambiguation handles both specific and general cases: for instance, only the locative case is retained in phrases like *v Ukrajini* ('in Ukraine.LOC'), and vocative forms are discarded after prepositions. Third, a statistical module draws on data from the manually disambiguated BrUK corpus (Starko and Rysin, 2023), using word form frequencies and morphological tags with contextual information (preceding and following token) to select the most probable analysis.[3] This hybrid approach shows promise but has not yet achieved complete reliability.

The corpus includes partial semantic annotation (Starko, 2021), which is stored in VESUM alongside morphological tags, and is thus assigned during the same tagging process. The annotation currently covers approximately 3,000 most frequent lemmas, as well as lemmas belonging to selected semantic groups, and continues to expand. A faceted approach is employed, allowing flexible tag com-

[3] https://github.com/brown-uk/nlp_uk/blob/master/doc/disambig.md

binations: lemmas receive one or more semantic features drawn from six major categories (concrete nouns, abstract nouns, proper nouns, adjectives, adverbs, and verbs), with separate tagsets developed for each category. For instance, the noun *ultras* receives the tags `conc:hum:group` (concrete noun, human, group). Some tags of a semantic nature — such as `prop` (proper noun) and its subtypes (e.g., `prop:fname` for first names, `prop:geo` for geographical names) — are incorporated into the morphological tag and assigned via the *tag* attribute, whereas the remaining semantic annotation is stored separately and searchable via the *semtag* attribute.

GRAC uses vertical files optimised for NoSketch Engine, where tokens carry positional attributes and structural metadata is encoded as attributes of the `<doc>` element. The compilation process consists of two stages: first, the TagText XML format is enriched with metadata and converted into a vertical file using XSLT; then, the resulting file is compiled into NoSketch Engine's indexed format using its CLI toolkit.

Users search via NoSketch Engine (Rychlý, 2007; Kilgarriff et al., 2014) using word forms, lemmas, tags, semtags, and complex CQL queries, with the ability to create random samples and obtain statistical information. Rich metadata enables fine-grained, multidimensional search queries.

The corpus is updated at least once a year. Before each update, metadata are validated and texts undergo preprocessing. Version-specific changes are documented on the project website.[4]

3. Challenges and Current Limitations

The corpus faces structural imbalances. Addressing representational gaps remains a long-term priority. Rather than aiming for a strictly balanced corpus in terms of historical versus contemporary coverage (which is unachievable given the uneven availability of sources) our goal is to ensure adequate representation across text types and regions. Contemporary online texts will likely continue to grow faster than historical collections, but targeted efforts to expand underrepresented text types will continue.

While morphological analysis handles standard contemporary Ukrainian effectively, several issues persist. Disambiguation accuracy requires continued improvement. Processing texts with non-standard orthography, particularly pre-standardization materials and regional variants, remains difficult despite specialized tools.

Mass-generated text poses a challenge for future corpus expansion. A related issue has already arisen with machine-translated content: many online media before 2014 published parallel Ukrainian and Russian versions, where the Ukrainian text may have been automatically translated, with or without post-editing. Since the origin of such texts could not be reliably determined, online media publishing two language versions were not included in the corpus. We are aware of the analogous risk posed by AI-generated content and are considering ways to address it, though no fully reliable detection method is currently available.

4. Future Plans

For GRAC v.20, we plan to implement syntactic dependency annotation using UDPipe2 (Milan Straka (2016–)). This will add a crucial layer to our annotation architecture, enabling more precise searches for grammatical phenomena and syntactic constructions. UD annotation should also improve disambiguation accuracy.

We continue working on corpus expansion with both historical and contemporary texts. Addressing the structural imbalances described above remains a long-term priority.

Despite ongoing challenges, GRAC has established itself as essential infrastructure for Ukrainian linguistics. With multi-level annotation, and growing derivative projects, it continues to serve research, education, and language technology.

5. Acknowledgements

We thank the reviewers for their valuable comments, which helped improve this paper. We are grateful to the GRAC team — Sergey Yarygin, Ruprecht von Waldenfels, Andriy Rysin, Vasyl Starko, Arsenii Lukashevskyi, and all others who have contributed to its development. We also acknowledge the support of the University of Jena and IPI PAN.

6. Bibliographical References

Yurii Chemerys, Olesia Nakhlik, Andriy Rysin, and Maria Shvedova. 2023. Normalization of a historic Western Ukrainian orthographic system Zhelekhivka in the Ukrainian language reference corpus (GRAC). In *Proceedings of the IEEE 18th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine.

[4] https://uacorpus.org/en/informaciya-pro-grak/versiyi-korpusu

CLARIN ERIC. 2025. PressMint: Interoperable corpora of historical newspapers. Accessed: 2026-02-28.

Olha Kanishcheva, Maria Shvedova, Liudmyla Dyka, and Kristina Husenko. 2026. Study of language identification task on the token level for Ukrainian-Russian code-switching dataset. *Northern European Journal of Language Technology*, 12(1).

Witold Kieraś, Łukasz Kobyliński, Dorota Komosińska, Michał Rudolf, Maria Shvedova, and Anna Zwierzchowska. 2025. PAWUK: Extensive annotated web corpus of Ukrainian. In *Computational Science – ICCS 2025*, volume 15904 of *Lecture Notes in Computer Science*, Cham. Springer.

Adam Kilgarriff et al. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Pavel Rychlý. 2007. Manatee/Bonito-a modular corpus manager. In *RASLAN*, pages 65–70.

Maria Shvedova, Arsenii Lukashevskyi, and Andriy Rysin. 2025. Developing a Universal Dependencies Treebank for Ukrainian parliamentary speech. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 55–63, Vienna, Austria. ACL.

Maria Shvedova, Nataliia Prydvorova, and Ilona Skibina. 2022. Normalization of early modern Ukrainian in GRAC: the case of Lesia Ukrainka's works. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022)*, Gliwice, Poland.

Maria Shvedova, Andriy Rysin, and Vasyl Starko. 2021. Handling of nonstandard spelling in GRAC. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, pages 105–108, Lviv, Ukraine.

Maria Shvedova and Ruprecht von Waldenfels. 2021. Regional annotation within GRAC, a large reference corpus of Ukrainian: Issues and challenges. In *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*, pages 32–45, Kharkiv, Ukraine.

Vasyl Starko. 2021. Implementing semantic annotation in a Ukrainian corpus. In *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*, pages 435–447, Kharkiv, Ukraine.

Vasyl Starko and Andriy Rysin. 2022. VESUM: A large morphological dictionary of Ukrainian as a dynamic tool. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022)*, pages 71–80, Gliwice, Poland.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics. DOI: 10.18653/v1/2023.unlp-1.11.

Vasyl Starko, Andriy Rysin, and Maria Shvedova. 2021. Ukrainian text preprocessing in GRAC. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 2, pages 101–104, Lviv, Ukraine.

7. Language Resource References

Arsenii Lukashevskyi, Kyrylo Zakharov, Maria Shvedova. 2025. *Rada_Trees: A Syntactically Annotated Corpus of Ukrainian Parliament Transcripts (1990–2024)*. Hugging Face Datasets.

Maria Shvedova, Ruprecht von Waldenfels, Sergey Yarygin, Andriy Rysin, Vasyl Starko, Tymofij Nikolajenko, Arsenii Lukashevskyi et al. 2017–. *GRAC: General Regionally Annotated Corpus of Ukrainian*.

Viktoriia Maslij (Kalashnyk) and Maria Shvedova. 2025. *ParaFarm: English-Ukrainian Multiple-Translation Corpus (1.1)*. Zenodo.

Milan Straka, Jana Straková, Jan Hajič. 2016–. *UDPipe Web Service (LINDAT/CLARIAH-CZ): Trainable Pipeline for Tokenization, Tagging, Lemmatization and Parsing*. LINDAT/CLARIAH-CZ, Institute of Formal and Applied Linguistics, Charles University.

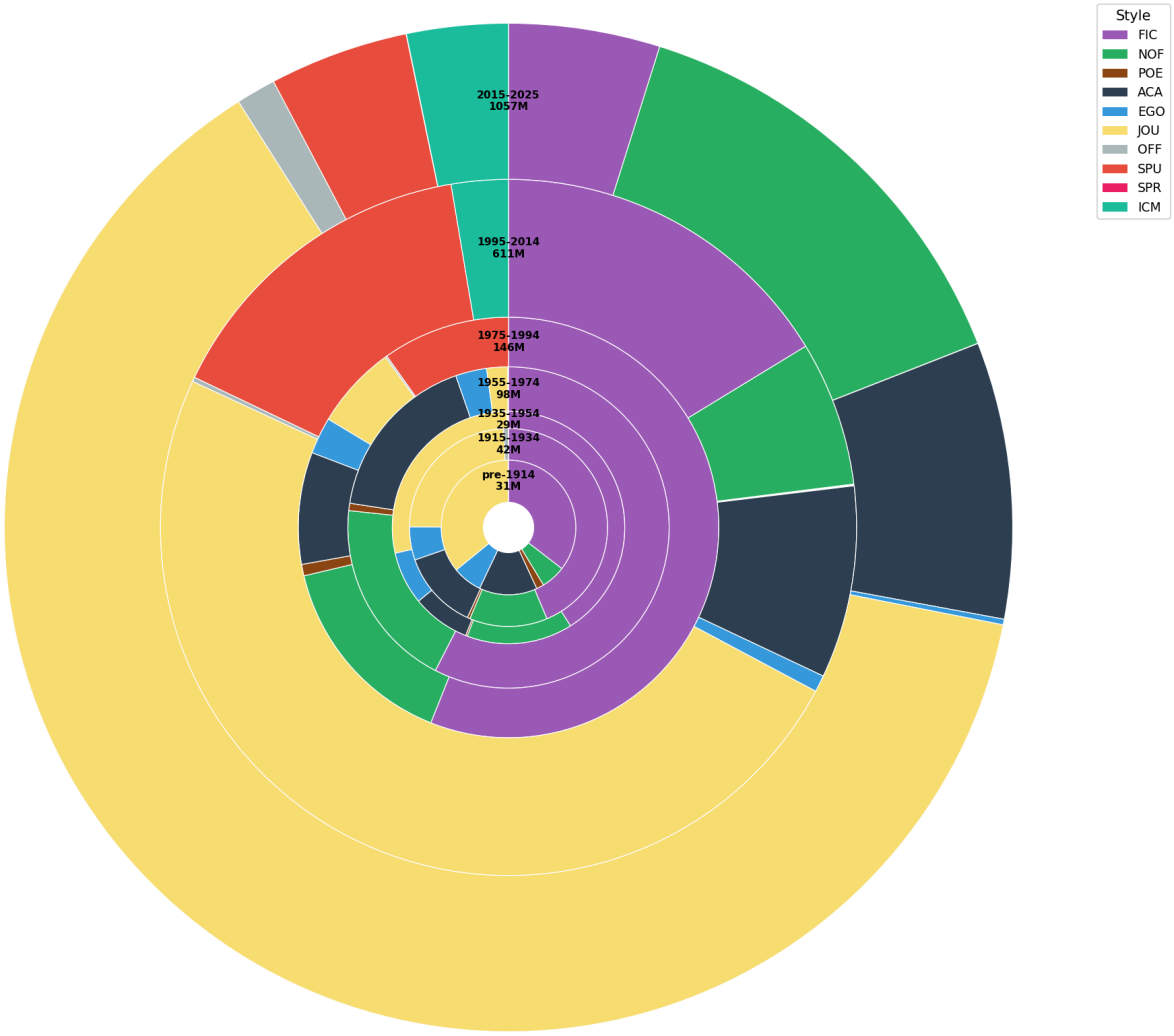
Olha Kanishcheva, Maria Shvedova, Liudmyla Dyka, Kristina Husenko. 2025. *ParlMix-UA-RU: Ukrainian Parliamentary Code-Mixing Dataset*. Zenodo.

Maria Shvedova and Arsenii Lukashevskyi. 2023–. *ParaRook: Parallel Corpora Based on GRAC*.

Maria Shvedova and Arsenii Lukashevskyi. 2024. *PluG: Corpus of Old Ukrainian Texts Based on GRAC*.

Maria Shvedova and Arsenii Lukashevskyi. 2024-2025. *UD_Ukrainian_ParlaMint*.

A. Distribution of functional styles across periods in GRAC.v.19 (ring area proportional to token count)



Style	pre-1914	1915-1934	1935-1954	1955-1974	1975-1994	1995-2014	2015-2025	**Grand Total**
FIC	11.0M (35%)	18.3M (44%)	12.1M (41%)	56.2M (58%)	81.9M (56%)	99.6M (16%)	51.2M (5%)	**330.2M (16%)**
NOF	1.8M (6%)	5.3M (13%)	4.3M (15%)	18.7M (19%)	22.3M (15%)	41.1M (7%)	150.0M (14%)	**243.6M (12%)**
POE	559.8K (2%)	185.5K (0%)	71.6K (0%)	723.8K (1%)	1.3M (1%)	506.8K (0%)	102.0K (0%)	**3.4M (0%)**
ACA	4.3M (14%)	5.4M (13%)	2.4M (8%)	16.9M (17%)	12.6M (9%)	54.6M (9%)	93.6M (9%)	**189.7M (9%)**
EGO	2.2M (7%)	2.3M (5%)	2.2M (7%)	3.1M (3%)	4.1M (3%)	4.7M (1%)	1.9M (0%)	**20.5M (1%)**
JOU	11.1M (36%)	10.1M (24%)	8.4M (28%)	2.1M (2%)	9.4M (6%)	300.1M (49%)	664.7M (63%)	**1005.8M (50%)**
OFF	15.5K (0%)	322.6K (1%)	26.6K (0%)	39.8K (0%)	200.1K (0%)	1.3M (0%)	13.5M (1%)	**15.3M (1%)**
SPU	16.8K (0%)	13.5K (0%)	41.4K (0%)	44.7K (0%)	14.4M (10%)	93.3M (15%)	47.1M (4%)	**154.8M (8%)**
SPR	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	122.7K (0%)	**122.7K (0%)**
ICM	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	16.2M (3%)	34.3M (3%)	**50.5M (3%)**
Total	**31M**	**42M**	**29M**	**98M**	**146M**	**611M**	**1057M**	**2014M (100%)**

Recent developments of the Bulgarian National Corpus

Svetla Koeva, Ivelina Stoyanova
Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
{svetla,iva}@dcl.bas.bg

Abstract

We present recent developments in the Bulgarian National Corpus, including data collection from various sources, cleaning of diverse datasets, enrichment with multimodal data, and extensive metadata, which resulted in the development of IfGPT, a large BulNC-based dataset. Typical methods for distributing the BulNC-based dataset are briefly described, with emphasis on effective searching within the metadata stored in a graph database.

1. Introduction

Over the past ten years, the Bulgarian National Corpus (BulNC) has undergone several developments to provide broader coverage of linguistic data and greater applicability to various NLP tasks. Since its establishment in 2009, the key features of BulNC have included **diversity of data** in terms of registers, domains, time periods, authors, and more; **multilinguality**; **extensive metadata** description; and **linguistic integrity**.

The significant development of BulNC, together with other national corpora in recent years, has been driven by the availability of large amounts of accessible data and new technologies for data collection, visualisation, and extraction of language facts and dependencies. Key areas of progress include the compilation and use of large volumes of multilingual and, to some extent, multimodal data; moving beyond simple corpus search and analysis to offer customised functions for linguistic analysis, such as defining words, tracking usage, detecting semantic shifts, and creating examples; serving as clean data for LLM pre-training and fine-tuning; and being analysed with LLMs.

The architectures of certain corpus management platforms enable the simultaneous processing of texts containing billions of words in many languages. For example, Sketch Engine provides access to over eight hundred corpora in more than one hundred languages, and allows complex linguistic queries and services (Kilgarriff et al., 2014).[1] English-Corpora.org is a collection of large corpora of English and its varieties, several of which contain billions of words (Davies, 2025).[2] The Czech National Corpus[3] provides access to written, spoken, parallel, and diachronic corpora comprising several billion words, which can be queried via the KonText interface (Machálek, 2020). The German Reference Corpus DeReKo, the largest corpus of written German, contains more than 60 billion words,[4] and is accessible through the KorAP corpus analysis platform (Diewald et al., 2016).

Recently, LLMs have been integrated into corpus query tools such as AntConc (Anthony, 2024), allowing identification of missing information and suggesting corrections for inconsistencies in dictionary drafts.

In addition to providing public access to the BulNC data for linguistic research, over the past ten years our efforts have focused on expanding BulNC with diverse and linguistically clean data suitable for NLP research and, more recently, for pre-training and fine-tuning LLMs. This has resulted in a shift in dataset accessibility, enabling the extraction of subcorpora for specific tasks based on extensive metadata.

These efforts have led to the development of the large **BulNC-based dataset** within the project *IfGPT: Infrastructure for Fine-tuning Pre-trained Large Language Models*[5] (also called the **IfGPT dataset**), with a special focus on the efficient management of large text data.

Alongside the expansion of textual data, we aim to provide more diverse data in terms of multilinguality (parallel corpora), various levels of annotation (including aligned corpora), and multimodal corpora suitable for a wide range of NLP and AI applications.

2. IfGPT, a large BulNC-based dataset

The components of **IfGPT, a BulNC-based dataset**, can be categorised into three main groups according to text type, composition, and potential uses: 1) collections of texts that have already been created, processed, and are available (BulNC belongs here); 2) other existing datasets of Bulgarian

[1] https://www.sketchengine.eu/
[2] https://www.english-corpora.org/
[3] https://www.korpus.cz/
[4] https://korap.ids-mannheim.de/
[5] https://ifgpt.dcl.bas.bg/en/

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 71–75. Palma, Mallorca. May 11, 2026.

texts that need to be reviewed, downloaded, and, if necessary, have their text and metadata formats converted to those of the IfGPT dataset; 3) compilation of new datasets through targeted crawling and processing of identified texts for filtering, cleaning, deduplication, and addition of metadata.

The new additions to IfGPT were collected, cleaned, and processed mainly within projects funded at national or European level, such as:

- **CEF Automated Translation for the EU Council Presidency**,[6] which involved collecting a large amount of parallel data and terminological resources in Bulgarian and English to train machine translation systems, focusing on official communication and translation challenges during the EU presidency.
- **Multilingual Resources for CEF.AT in the Legal Domain** (**MARCELL**),[7] which collected national legislative texts in seven languages – Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, and Slovenian – and annotated each subcorpus with morphosyntax, dependency structure, named entities, and IATE/EuroVoc terminology.
- **Curated Multilingual Language Resources for CEF AT** (**CURLICAT**),[8] which collected, cleaned, anonymised, and annotated licence-free texts in the same seven languages, producing over 14 million sentences across the health, culture, science, and government domains, with a harmonised metadata schema designed for neural machine translation.
- **Infrastructure for Fine-tuning Pre-trained Large Language Models** (**IfGPT**), which aims, among other tasks, to collect, filter, anonymise, and deduplicate large, diverse, high-quality text data for fine-tuning pre-trained LLMs for Bulgarian.

Further extensions of the dataset include newly collected and processed texts from various time periods. Older texts, such as news articles, periodicals, and books published before 1990, are also collected and processed using OCR. Table 1 shows the most important parts of the IfGPT dataset.

Source	# texts	# tokens	Licence
MARCELL	25K	45M	PD
CURLICAT	113K	35M	CC
BulNC Admin	17K	79M	PD
BulNC Wikipedia	89K	41M	CC/GNU
BulNC Subtitles	146K	27M	OPUS
BG News	2,116K	601M	various
EN News	5,961K	3,324M	various
BG internet	66K	289M	various
EN internet	45K	8,144M	various
News up to 1990	5,544K	270,52M	various
Periodicals up to 1990	25K	30M	various
New periodicals	4,119K	4,378M	various
Books	22K	630M	various

Table 1: Current structure (March 2026). Licences: PD – public domain, CC – Creative Commons (various), GNU – GNU free license, various – other (including restrictive) licences.

3. Multimodal Data

The BulNC-based dataset is extended with multimodal data from the **Multilingual Image Corpus (MIC21)** (Koeva et al., 2022). MIC21 provides pixel-level annotations for over 203,000 objects in more than 21,000 images, covering 730 object classes across four thematic domains.[9]

The images are carefully selected to ensure high-quality, copyright-free content from thematically related domains (Sport, Transport, Art, and Security), comprising 130 related subdomains, and are supplied with available metadata. Annotation is performed by drawing or correcting automatically generated polygons using the Detectron2 model (Wu et al., 2019)), from which bounding boxes are then generated automatically. This enables wide application of the dataset in various computer vision tasks: image classification, recognition and classification of single objects in an image, or of all object instances in an image (semantic segmentation). An example of an image from the domain **Art** and the subdomain **Violinist**, with three annotated objects (*violinist*, *violin*, and *bow*), is shown in Figure 1.

The classes for object annotation are organised in an Ontology of Visual Objects (Koeva, 2022), which offers options for extracting relationships between objects in images, constructing diverse datasets with varying levels of object class granularity, and compiling suitable sets of images illustrating different thematic domains. Some classes and relations are inherited from WordNet. Additional classes and relations are included in the ontology if they are not present in WordNet; for example, ***Bowler*** *wears* ***Bowling shoes***. Object classes are linked to certain metadata values of the BulNC-based dataset, thereby relating visual content to the textual dataset.

The object labels from the Ontology are linked to their synonyms, definitions, and usage examples in 25 languages. The selection of languages was

[6] https://tilde.ai/machine-translation/
[7] https://marcell-project.eu
[8] https://curlicat.eu
[9] https://dcl.bas.bg/en/projects_list/mic21/

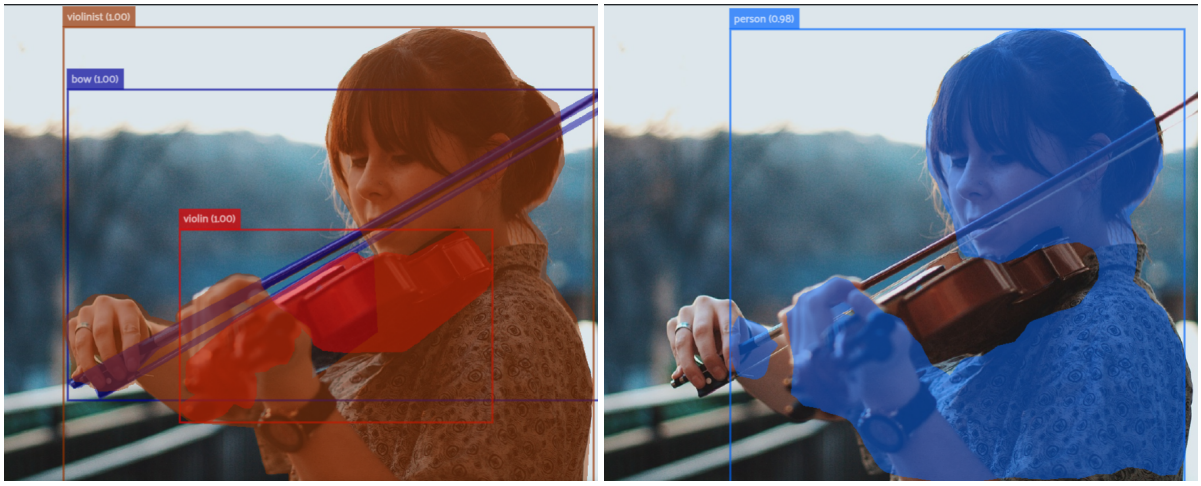


Figure 1: Image with masks, bounding boxes and labels: on the left – MIC21 manually annotated; on the right – automatically annotated using the Detectron2 model.

based on the availability of wordnets in various languages in the Extended Open Multilingual Wordnet (Bond and Foster, 2013). Where WordNet translations are unavailable, additional sources of translations are used, mainly BabelNet[10] and machine translation. The labels of objects whose concepts are not present in WordNet have been translated by experts only into English and Bulgarian. The multilingual layer makes the dataset suitable for artificial intelligence applications such as multilingual image captioning, question answering, and machine translation of multimodal content.

Recently, images have been provided with short narrative descriptions that explain the relationships between the depicted objects.

4. IfGPT dataset processing pipeline

For the expansion of the BulNC, the integration of resources developed through collaborative international projects, and the preparation of data tailored for language technologies and LLMs, the following components of the **IfGPT dataset processing pipeline** have been developed:

- **File handling module** for managing files in appropriate formats for text and metadata (e.g., plain text, JSONL, CSV), using the adopted metadata schema.
- **Dataset quality maintenance module** providing functions for string manipulation, data cleaning, and error handling to support data quality assurance through text deduplication, identification and labelling of personally identifiable information, and detection of potential bias.
- **Metadata extraction module** for obtaining metadata from the document source and content, and providing appropriate metadata descriptions.
- **Annotation module** for introducing traditional linguistic annotation in CoNLL-U Plus format (Koeva et al., 2020).
- **Dataset construction module** for creating subdatasets for specific purposes based on extensive metadata.
- **Search module** providing an online interface for browsing metadata values for selection (Koeva et al., 2025). Its output is either a newly constructed subdataset or a selection of links for downloading relevant parts of the subdataset.

5. Metadata and ways of distribution

The metadata is designed for searching and retrieving information to support various research and applications, and therefore has a complex graph-based structure of related categories (Koeva et al., 2016). The metadata of the IfGPT dataset originates from the BulNC and is harmonised with the metadata of newer multilingual corpora such as MARCELL and CURLICAT.

The metadata includes 15 mandatory categories covering technical details (such as identifier), source description (source URL, licence), and document statistics (number of words, sentences), as well as 9 optional categories describing features of the document (such as author, style). The metadata is managed using the graph database Neo4J,[11] which is designed to handle large vol-

[10] https://babelnet.org/

[11] https://neo4j.com/

umes of interconnected data efficiently and maintains performance under complex queries using the Cypher query language (Francis et al., 2018). The graph database effectively models relations between metadata values (e.g. `WRITTEN_BY` for authorship, `LICENCED_WITH` for licensing, `BELONGS_TO` for domain classification) and allows efficient access to the metadata and extraction of different subsets according to users' needs.

The Bulgarian National Corpus offers a customised web interface for searching the corpus, building concordances, and extracting examples (Koeva et al., 2012, 100-101).[12] The search system supports complex linguistic queries involving different levels of annotation (POS, morphosyntactic features, semantic relations) combined in various ways.

Parts of the BulNC and the extended IfGPT dataset with open licences are available for direct download, while some parts are subject to copyright restrictions. The latter may be used to compile subsets for specific users and tasks, but cannot be redistributed directly.

The IfGPT Dataset search interface[13] allows users to browse and filter the large collection of clean, deduplicated Bulgarian text documents by several criteria: type of licence, domain, time period, and keywords. The user can export metadata, links, or raw texts. It is aimed for dataset compilation and extraction of data for use in NLP applications and LLM fine-tuning.

6. Conclusion and future work

In summary, the IfGPT dataset, based on BulNC, is designed to be as large as possible while incorporating rich metadata to support efficient search and retrieval of relevant data for research and practical applications, including the pre-training and fine-tuning of large language models. Future development of IfGPT will include improvements in data distribution and curation methods, such as integrating large language models to identify missing information, resolve inconsistencies, and enhance the overall dataset compilation process.

Extremely large text collections have been developed by crawling internet data. For example, Common Crawl contains petabytes of data, including Bulgarian. It includes raw web page data, metadata extracts, and text extracts (Common Crawl Foundation, 2025). Many LLM datasets have been created based on it, such as mC4, OSCAR, and CulturaX. One of the largest, CulturaX, is a multilingual dataset with 6.3 trillion tokens in 167 languages, including Bulgarian. The dataset undergoes extensive cleaning and deduplication. The HPLT Corpus (Burchell et al., 2025) contains monolingual corpora covering 193 languages, including Bulgarian, and approximately 8 trillion tokens. Parallel corpora from monolingual data for 50 languages paired with English were derived, containing over 380 million sentence pairs. The corpus was extracted from 4.5 petabytes of Internet Archive and Common Crawl data.

Ensuring the quality of large datasets is a critical prerequisite for efficient and reliable training of LLMs, as noise, redundancy, and malformed content directly degrade model performance and introduce systematic biases (Kreutzer et al., 2022). For Bulgarian, a morphologically rich and low-resource language, ensuring high data quality is challenging due to limited tools for efficient identification of text matches and near matches, inconsistent orthography, incomplete sentences, and texts produced using machine translation. Deduplication is especially important at scale, particularly for near-duplicate documents, which can artificially increase corpus size and cause language models to over-represent certain linguistic patterns.

Developing robust automatic methods for these tasks for Bulgarian is difficult, as pipelines implemented for high-resource languages are not always suitable or straightforward to adapt, while building language-specific solutions requires significant resources, including manual annotation and evaluation. Thus, developing efficient and scalable methods for quality checks and improvement of large datasets in Bulgarian remains an open research challenge.

7. Acknowledgments

The present study is carried out within the project Infrastructure for Fine-tuning Pretrained Large Language Models, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Laurence Anthony. 2024. Antconc (version 4.3.0).

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komu-

[12] https://search.dcl.bas.bg/

[13] https://ifgpt.dcl.bas.bg/ifgpt-dataset/

lainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Common Crawl Foundation. 2025. Common crawl web corpus.

Mark Davies. 2025. English-corpora.org.

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP architecture – diving in the deep sea of corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3586–3591, Portořož, Slovenia. ELRA.

Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An Evolving Query Language for Property Graphs. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1433–1445, New York, NY, USA. Association for Computing Machinery.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1(1):7–36.

Svetla Koeva. 2022. Ontology of visual objects. In *Proceedings of the Fifth International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 120–129, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. ELRA.

Svetla Koeva, Ivelina Stoyanova, and Jordan Kralev. 2022. Multilingual image corpus – towards a multimodal and multilingual dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1509–1518, Marseille, France. ELRA.

Svetla Koeva, Ivelina Stoyanova, and Jordan Kralev. 2025. IfGPT: A dataset in Bulgarian for large language models. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 65–75, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 1(1):65–110.

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva, and Tsvetana Dimitrova. 2016. Metadata extraction, representation and management within the Bulgarian National Corpus. In *Priceedings of the 4th Workshop on Challenges in the Management of Large Corpora*, pages 33–39. ELDA.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 7003–7008, Marseille, France. ELRA.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. GitHub.

British National Corpus 1994 to 2026

Martin Wynne, Megan Bushnell

University of Oxford
Faculty of Linguistics, Philology and Phonetics, Oxford, UK
{martin.wynne, megan.bushnell}@ling-phil.ox.ac.uk

Abstract

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. It is one of the first generation of monolingual, synchronic, general, representative corpora of its size, and led the way for other national corpora. It was created by a consortium of academic partners and publishers, with funding from the Department of Trade and Industry in the UK. This posters reflects on a number of lessons learning in more than thirty years, in terms of corpus representativeness, modes of access to the corpus, licensing, and managing the transition from a contemporary synchronic corpus to a historical corpus.

Keywords: Corpus linguistics, linguistic corpus, corpus construction, licensing

1. Extended abstract

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. It is one of the first generation of monolingual, synchronic, general, representative corpora of its size, and led the way for other national corpora. It was created by a consortium of academic partners and publishers, with funding from the Department of Trade and Industry in the UK[1].

The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

The corpus is encoded according to the Guidelines of the Text Encoding Initiative (TEI) to represent both the output from CLAWS (automatic part-of-speech tagger)[2] and a variety of other structural properties of texts (e.g. headings, paragraphs, lists etc.). Full classification, contextual and bibliographic information is also included with each text in the form of a TEI-conformant header. The corpus was an important landmark in the adoption of the TEI guidelines for a linguistic corpus. The creation of the corpus and its structural markup and annotation are fully documented in the User Reference Guide[3].

Work on building the corpus began in 1991, and was completed in 1994. No new texts have been added after the completion of the project but the corpus was slightly revised for copyright reasons prior to the release of the second edition BNC World (2001) and the third edition BNC XML Edition (2007). The part-of-speech tagging was revised and improved in the BNC Tag Enhancement project 1995-1996 at Lancaster University. The two-million-word BNC Sampler was manually annotated, then used to train the CLAWS part-of-speech tagger before automatically re-tagging the remaining c.98 million words of the corpus. The morphosyntactic annotations assigned at this time remain in all officially released versions of the corpus.

The BNC was originally made available on CD-ROM bundled with the Sara software for analysis and exploration of the text and tagging, with users paying a fee for media and administrative costs. Later, this was done on DVD with the Xaira (XML-aware) software, and in 2014 the BNC became available for download for free from the Oxford Text Archive, the CLARIN-UK repository. The Oxford Text Archive has been collecting corpora and other electronic texts and datasets since 1976, and celebrates its fiftieth anniversary in 2026.

A major project in the 2000s located a large proportion of the audio files on which the spoken corpus is based, aligned them with the text, and made them available from a streaming server[4]. A version of BNCWeb[5] makes use of this facility to offer the audio for concordance lines.

[1]https://www.natcorp.ox.ac.uk/corpus/creating.xml
[2]https://ucrel.lancs.ac.uk/claws/
[3]https://www.natcorp.ox.ac.uk/docs/URG/
[4]https://www.phon.ox.ac.uk/AudioBNC

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 76–77. Palma, Mallorca. May 11, 2026.

The licence for the BNC, agreed with the copyright owners of the materials included, only allows distribution on CD-ROM by Oxford University Computing Services on behalf of the BNC Consortium. This has been reinterpreted to allow online download, but only from the University of Oxford. However, the licence has been interpreted in such a way as to allow online corpus platforms to host the corpus and allow analysis and exploration, but not download of whole texts or the whole corpus. These platforms, including BNCWeb, English-Corpora.org and Sketch Engine, have been intensively used for many years. The BNC Licence is unusual in that it expressively allows and encourages commercial use of the corpus.

Researchers at Lancaster University created a comparable corpus BNC2014[6], a synchronic corpus of present-day English from a period 20 years after the original BNC. To facilitate identification and comparison of comparable corpora, the original BNC has been rebranded as BNC1994. One important lesson learned from the long history of the BNC, is that it is necessary to transition from branding a corpus as a snapshot of present-day language to a historical corpus. The corpus now also represents an important source of human language produced by native speakers from just before the internet age and the arrival of computer-mediated modes, and also from a time before the pollution of language data with computer-generated language.

2. Bibliographical References

BNC Consortium, British National Corpus 1994, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2554

Coleman, J., Baghai-Ravary, L., Pybus, J., and Grau, S. (2012). Audio BNC: the audio edition of the Spoken British National Corpus. Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/AudioBNC

Garside, R. (1996). The robust tagging of unrestricted text: the BNC experience. In J. Thomas and M. Short (Eds), Using corpora for language research: Studies in the Honour of Geoffrey Leech Longman, London, pp 167-180.

Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Kyoto, Japan, pp. 622-628.

Burnage, G. and Dunlop, D. (1993) Encoding the British National Corpus. In J. Aarts, P. de Haan and N. Oostdijk (Eds), English language corpora: design, analysis and exploitation. Amsterdam: Rodopi, pp. 79-95.

3. Language Resource References

BNC Consortium, British National Corpus 1994, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2554

BNC Consortium, British National Corpus 1994 Sampler, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2552

[5]http://bncweb.lancs.ac.uk/
[6]http://corpora.lancs.ac.uk/bnc2014/

The Corpus of Contemporary Polish: 2011-2020 Decade and Beyond

Witold Kieraś, Małgorzata Marciniak, Katarzyna Krasnowska-Kieraś, Marcin Woliński
Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
{w.kieras, m.marciniak, k.krasnowska-kieras,m.wolinski}@ipipan.waw.pl

Abstract

The aim of this poster is to present the Contemporary Corpus of Polish (KWJP), a new reference resource spanning the period of 2011–2020. The KWJP complements the now discontinued National Corpus of Polish project (NKJP, Przepiórkowski et al. 2012) by providing up-to-date linguistic data. It comprises a 100M-token balanced sub-corpus alongside a larger 1.5B-token unbalanced (opportunistic) component, consisting of books and periodicals not included in the balanced part. While the corpus contains almost exclusively copyrighted material and is therefore accessible only via a web-based search engine, a representative 0.5M-token sample has been published as open-source data. Details of the resource description that fall beyond the scope of this abstract can be found in a paper accepted for the main LREC 2026 conference (Kieraś et al., 2026).

While the KWJP was conceived as a successor to the NKJP for the subsequent decade, it differs from its predecessor in several key respects. A primary distinction lies in its temporal scope: whereas the NKJP aimed to represent written Polish from the early 20th century onwards, the KWJP focuses exclusively on a single decade of contemporary texts. Given the emergence of numerous specialized resources—such as the Corpus of Parliamentary Discourse (Ogrodniczuk, 2018), the MoncoPL web monitoring corpus (Pęzik, 2020), and various spoken language corpora (Pęzik, 2015)—the necessity for a general reference corpus to cover every possible genre has diminished. Consequently, the KWJP focuses strictly on edited texts, namely books (both fiction and non-fiction) and a broad selection of national and regional periodicals.

Consequently, the KWJP's text type classification and the proportions represented in the corpus have been significantly simplified, now consisting of three categories: fiction, non-fiction, and journalism. Fiction (30% of the balanced corpus) primarily comprises literary books (novels and short story collections) across various genres, along with a limited selection of literary periodicals that publish predominantly short stories. The non-fiction category (35%) encompasses a broad range of texts, including journalistic books, diaries, biographies, travel guides, popular science, and scholarly works, as well as official documents—all of which were distributed across several distinct labels in the NKJP. Unlike the NKJP, thematic magazines are also classified as non-fiction. Finally, the journalism genre (35%) consists exclusively of traditional news and public affairs press, mainly national and regional daily and weekly newspapers, supplemented by a selection of monthly, bi-monthly, and annual periodicals.

Distribution channels are categorized into two primary types: books and press. Only a negligible portion of the texts (0.3%) is assigned to the internet channel, representing a sample of court rulings from various judicial instances—a specific type of official document. In all other cases, classification into the book or press channel follows standard library identification schemes, specifically ISBNs for books and ISSNs for the press, even for publications existing solely in electronic format. Books comprise approximately 55% of the balanced corpus, while the press accounts for the remaining 45% (daily newspapers: 19%; weekly magazines: 12%; monthly periodicals: 9.5%; other: 5.5%).

The KWJP features rich, multi-layer annotation, generally adhering to the NKJP annotation scheme. Rules for segmentation (tokenization) are followed directly. Regarding morphosyntactic tagging, the tagset has been aligned with the latest version of the Morfeusz morphological analyzer for Polish (Kieraś and Woliński, 2017; Woliński, 2014). Additionally, the KWJP introduces two new annotation layers: named entities (NE) and syntactic structures. Named entity annotation strictly follows the schema used in the one-million-word manually annotated subcorpus of the NKJP (NKJP1M, Przepiórkowski et al. 2012). As opposed to the NKJP, the automatic NE layer in the KWJP covers the entire resource. The syntactic layer is entirely new compared to the NKJP and comprises hybrid tree structures that combine both dependency and constituency relations (Krasnowska-Kieraś and Woliński, 2024). All annotation layers are accessi-

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 78–79. Palma, Mallorca. May 11, 2026.

ble (to a certain extent) via corpus queries (CQL).

The KWJP project aims to establish a team and infrastructure for the long-term development of the resource. The first update is scheduled for 2026 and will introduce a sub-corpus covering the 2021–2025 period. We intend to maintain a regular five-year update cycle, similar to the SYN corpora series developed by the Czech National Corpus team (Křen et al., 2016). The estimated minimum size of the update is 50 million tokens, keeping the same proportions as the original 2011–2020 corpus. Simultaneously, technical work is underway to provide a new, more efficient search engine that will support the development of web-based applications and the gathering of statistics for linguistic research.

1. Bibliographical References

Witold Kieraś, Małgorzata Marciniak, Marcin Woliński, Katarzyna Krasnowska-Kieraś, and Marek Łaziński. 2026. The Corpus of Contemporary Polish — a New Reference Corpus with Rich Syntactic Annotations. In *Proceedings of LREC 2026*, Palma, Spain.

Witold Kieraś and Marcin Woliński. 2017. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1):75–83.

Katarzyna Krasnowska-Kieraś and Marcin Woliński. 2024. Parsing headed constituencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12633–12643, Turin, Italy. ELRA and ICCL.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Jan Zasina. 2016. SYN2015: Representative corpus of contemporary written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2522–2528, Portorož, Slovenia. European Language Resources Association (ELRA).

Maciej Ogrodniczuk. 2018. Polish Parliamentary Corpus. In *Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, pages 15–19, Paris. European Language Resources Association (ELRA).

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Piotr Pęzik. 2015. Spokes – a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*, Linköping Electronic Conference Proceedings, pages 99–109. Linköping University Electronic Press, Linköpings universitet.

Piotr Pęzik. 2020. Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*, 7(7):133–150.

Marcin Woliński. 2014. Morfeusz reloaded. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. European Language Resources Association (ELRA).

Building the v4 of the Croatian National Corpus

Marko Tadić, Vanja Štefanec, Daša Farkaš

University of Zagreb Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
{marko.tadic, vstefane, dfarkas}@ffzg.unizg.hr

Abstract

It has been thirteen years since the release of the current version (v3) of the Croatian National Corpus (HNK). In terms of synchronicity in corpus linguistics, that many years may be considered quite some time. The preparatory phase for the composition of the new version of HNK (v4) has been going already for several years and in this paper we touch on several issues of concern. Apart of regular corpus parameters, e.g. text sources, text genres, coverage of language varieties, time span, we also discuss about metadata and linguistic annotation schemata. One of important technical prerequisites was the development of CorpRepo, a custom corpus data management system and file system, which enable us to do sustainable long-term maintenance of the data, and to produce newer versions of corpus more easily and more often. The selection of IPR-cleared data entails some restrictions and we give several examples of that kind of textual sources, but also discuss possible weaknesses of such approach to data selection. Regarding the linguistic annotation, the important shift is the decision to abandon the MulText East morphosyntacting descriptions and use solutions recommended by UD-initiative.

Keywords: Croatian National Corpus, corpus annotation, corpus data management

1. Introduction

It has been more than ten years since the release of the current version, version 3, of the Croatian National Corpus (HNK), in 2013 and seventeen years since the version 2.5 in 2009 (Tadić, 2009). In terms of synchronicity in corpus linguistics, that many years may be considered quite a long time since not just new texts appeared, but the distribution of types and genres as well as means and channels of text circulations in society also changed. The same goes for advances of annotation tools, since the HNK has not been re-annotated and no new annotation layers were added during all this time.

The preparatory phase for the composition of the new version of HNK (v4) has been going already for several years during which we have been considering various issues like text sources, text genres, coverage of language varieties, metadata, time span, structural and linguistic annotation schemata.

We've also been focusing on the development of the custom corpus data management system, which would enable us to do sustainable long-term maintenance of the data, and to produce newer versions more easily and more often.

In the times when sizes of large corpora exceed several dozens of billions of tokens, and LLMs are routinely used to generate textual content, it is difficult to find justification for composing a moderately-sized hand-picked national corpus. After all, corpus linguistics could be regarded as a data science in which “more the data, more the value”. We, however, argue that in fact it does make sense to embark on that endeavour because humanly curated large representative corpora should still be considered a gold standard in corpus linguistics.

The paper presents the topic of text types and related IPR issues in section 2. The section 3 is explaining the metadata approach while the section 4 describes the custom corpus data management system. In section 5 the plans for linguistic annotation are laid down and the paper ends with conclusion as section 6.

2. Text Types and IPR

It is planned that the new version of HNK will, unlike its predecessors, contain IPR-cleared texts and will be made freely available for academic purposes under a permissive license. This means that all texts will be either collected from publicly available sources or acquired from public institutions who are obliged by law to give access to textual data they are producing or managing in some way, for scientific purposes. On top of that we are in the process of negotiating the data-providing agreements with important Croatian publishers like Matica hrvatska and Lexicographic Institute Miroslav Krleža. Since they receive support from the Ministry of Culture and Media and/or Science for publishing their editions, these might be also partially available under permissive license. At this moment we can't predict the outcome of these negotiations and any projections on these text types and their size would be highly speculative. However, we will do our best to include as much fiction as possible and we already have some texts, that have been donated by the authors themselves.

This intention to use IPR-cleared data could in fact turn the expected well-balanced and

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 80–83. Palma, Mallorca. May 11, 2026.

representative corpus into a corpus that could be called an opportunistic one since it might not follow the balanced representation of different text types, genres, domains, etc. We are well aware of that possibility, but we expect that the impact and usability of freely available very large corpus can be more important than meticulously followed theoretically planned structure.

Text types that are IPR-cleared include, for example, papers from the Portal of Croatian scientific and professional journals[1] published under different versions of open access. This portal includes 572 scientific and professional journals with more than 322,000 full-text papers from all domains of science. These papers together with BA, MA and PhD theses in open access at Digital Academic Archives and Repositories[2] from different Croatian universities and from the National and University Library form the Croatian Scientific Corpus, which will be a part of the HNK v4. The estimated size of this subcorpus is more than 600 Mw.

Another example of IPR-cleared text types is the corpus of legal texts published in the official journal Narodne novine[3]. This includes texts of laws and lower-level legal documents of the Parliament, Government, regional authorities, Constitutional Court and Croatian Central Bank. Additional documents of local authorities are available in the digital form through the Central Catalogue of the Official Documents of the Republic of Croatia[4]. Texts from both sources were partially included in the Croatian MARCELL Legislative Suborpus (Váradi et al, 2020), which was 102 Mw in size at that time. However, we are planning to include also texts produced since that time until today.

Tentative list of text types and their approximate proportions for HNK v4, while the desired target size is at least 1Gw:

- newspapers and magazines 50%
- legislative and public texts 15%
- academic 15%
- literature 10%
- publicistics 5%
- mixed types 5%

3. Metadata

Given the fact that also the structure of corpus users has broadened since the current HNK v3 appeared and it nowadays includes researchers from all fields and branches of humanities as well as social sciences, special attention will be put on catering for their specific requirements. This includes a much richer document metadata description, as well as preserving and unifying general structures within a document (titles, headings, paragraphs, articles, etc.), which were rather inconsistent in HNK v3.

In the case of two mentioned sources for scientific text types (Hrčak and Dabar repositories), they use slightly different metadata schemata, so we have built a new (meta)data model in order to harmonize the description of their objects. After importing all the metadata in the database, we harvested the actual objects from the respective repositories.

Metadata describing published scientific papers contain, among others, information about the classification using the triple layered hierarchy in different scientific domains, fields and branches in accordance with the Croatian regulations on classification of domains of sciences and arts[5]. This information from metadata facilitates the generation of domain-specific subcorpora and research that will enable the comparison of texts between different scientific fields.

4. Data Management

In very large corpora data management represents a specific challenge, but we still wanted to have a sustainable data management that is flexible enough regarding the metadata control, harvesting of documents, extraction of raw text, etc. Since for corpora data management there are no universally applicable guidelines, we set our own requirements on data management at least for large systematized sources of data like institutional repositories: 1) save all metadata, 2) version-track all documents, 3) enable collaborative work, 4) provide work environment that ensures long-term sustainability of data. We developed CorpRepo, a custom database-driven software solution (web-application), which can control large number of git repositories. It enables us to: 1) ingest and parse document metadata, 2) harvest documents and store them on file-system, 3) perform necessary git repository actions (add, commit, push, pull), 4) perform automatic, semi-automatic and manual document editing, 5) calculate and re-calculate relevant statistics, 6) generate corpora based on various parameters.

4.1 Metadata ingestion and processing

Metadata is harvested through repositories' OAI-PMH interface while webapp parses metadata, extracts relevant data and stores them into database along with the full metadata record. In many cases the metadata records also contain document abstracts.

[1] https://hrcak.srce.hr/
[2] https://dabar.srce.hr/
[3] https://www.nn.hr/
[4] https://sredisnjikatalogrh.gov.hr/
[5] https://narodne-novine.nn.hr/clanci/sluzbeni/2024_01_3_69.html

4.2 Document harvesting

Webapp takes document URL (or some other permanent identifier) from the metadata record, downloads the document and saves it onto the file-system.

4.3 Text extraction

For text extraction we're using GROBID[6], an ML library for parsing and re-structuring raw documents into structured XML/TEI. Webapp takes the document from the file-system and sends it to GROBID API where text is extracted from the resulting XML/TEI and saved onto file-system, along with the original TEI

4.4 Repository management

In repository management webapp controls both, the file-system and git repositories. All changes in the file-system are version-tracked.

4.5 Data cleaning

Through the webapp user can perform semi-automatic or manual cleaning of the data. Accuracy of the text extraction process varies significantly based on the document layout and date of creation. Text extracted from PDF documents is extremely noisy, especially in Hrčak dataset.

5. Linguistic Annotation

The largest improvement will be made with the linguistic annotation. HNK v3 has been annotated only on morphosyntactic level and lacked the annotation of syntactic and semantic relations or named entities. Moreover, the annotation was performed using the Croatian MULTEXT-East tagset v4[7] (Erjavec, 2010), composed according to the specifications defined within the MULTEXT project (Dimitrova et al., 1998). With the introduction of Universal Dependencies (UD) initiative[8] (de Marneffe et al., 2021) and the inclination of the linguistic community towards the annotation standard it proposed and constantly develops, we decided to discontinue the use of MULTEXT-East tagset. We also believe that the basic UD tagset is much more legible to wider audience of users, less redundant, and much better documented. Also, given that the central tendency of UD is to provide a framework for consistent annotation of grammar across different human languages, we believe that this will increase the importance of the HNK v4 in the cross-linguistic research as well.

When it comes to the access to the corpus through a concordancing interface, we plan to maximally simplify creating even the very complex queries by creating various attributes during preprocessing. This primarily refers to annotating perifrastic verb forms, and performing traditional, non token-based lemmatization, which will enable even non experts in the field, to use the corpus in research, teaching, language-learning, or merely for information purposes.

6. Conclusion

We have presented the work-in-progress on the development of the v4 of the Croatian National Corpus (HNK v4) and several areas of interest to the main topic of this workshop. Once completed, HNK v4 will be searchable online through HR-CLARIN concordancer[9] as well as the Sketch Engine system (Kilgarriff et al., 2004). The expected date of release is the end of 2026 or beginning of 2027.

7. Acknowledgments

This research has been partially funded by the Ministry of Science, Education and Youth of the Republic of Croatia through the support to the HR-CLARIN consortium, a Croatian participation in the CLARIN ERIC.

8. Bibliographical References

de Marneffe, M.-C., Manning, C., Nivre, J. and Zeman, D. (2021). Universal Dependencies. In *Computational Linguistics* 47(2): 255--308.

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic and Tufiş, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of COLING-ACL 1998*, pages 315--319, Montreal, Canada. ACL.

Erjavec, T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2544--2547, Valletta, Malta, May. European Language Resource Association (ELRA).

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105--116, Lorient, France, July.

Ljubešić, N. and Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, pages 29--34, Florence, Italy, August.

Silić, J. (2006). *Funkcionalni stilovi hrvatskoga jezika.* Zagreb: Disput.

[6] https://github.com/kermitt2/grobid
[7] https://nl.ijs.si/ME/Vault/V4/msd/html/msd-hr.html
[8] https://universaldependencies.org/
[9] https://corpora.clarin.hr

Tadić, M. (2009). New version of the Croatian National Corpus. In Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.), *After Half a Century of Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 219--228.

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R. Krek, S., Repar, A., Rihtar, M. and Brank, J. (2020). The MARCELL Legislative Corpus. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 3761--3768, Marseille, France, May. European Language Resource Association (ELRA).

9. Language Resource References

Erjavec, Tomaž; et al., 2025, Multilingual comparable corpora of parliamentary debates ParlaMint 5.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/2004.

MARCELL Croatian Legislative Subcorpus. 2020. European Language Grid repository, https://live.european-language-grid.eu/catalogue/corpus/21358.

Tadić, Marko. 2014. Croatian National Corpus v3, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), http://hdl.handle.net/11372/LRT-233.

Managing Growth in a National Corpus: The Hungarian National Corpus 3.0 (MNSZ3)

Noémi Ligeti-Nagy[1], **Enikő Héja**[1], **Ágnes Bánfi**[1], **Flóra Földesi**[1], **Bence Sárossy**[1], **Boglárka Skrabák**[2], **Tamás Váradi**[1], **Gábor Prószéky**[1]
[1] ELTE Research Centre for Linguistics, Budapest, Hungary
[2] ELTE Faculty of Informatics, Budapest, Hungary
ligeti-nagy.noemi@nytud.hu, heja.eniko@nytud.hu, banfi.agnes@nytud.hu, foldesi.flora@nytud.hu, sarossy.bence@nytud.hu, skrabakbogi@gmail.com, varadi.tamas@nytud.hu, proszeky.gabor@nytud.hu

Abstract

The third generation of the Hungarian National Corpus (MNSZ3) aims to provide a large-scale, curated, and well-described corpus resource needed for the sustainable digital presence of Hungarian. Building on the domain structure and proportions of MNSZ2 (v2.0.5; 1.04 billion running words), the project targets a substantial increase in scale while also strengthening the coverage and metadata description of Hungarian language use outside Hungary. MNSZ3 retains the six traditional domains of the earlier corpus—press, fiction, scientific, official, personal, and transcribed spoken language—and is planned to reach approximately 10 billion tokens. This paper presents the motivation and design principles of the project, outlines the practical decisions and procedures used in data collection and cleaning, and discusses the annotation strategy developed for large-scale processing. In planning the linguistic analysis, we build on the complementary strengths of HuSpaCy and e-magyar: HuSpaCy provides the unified and efficient UD-oriented processing backbone, while e-magyar (emMorph) is preserved as an explicit additional layer for morphology and lemmatisation.

Keywords: Hungarian, national corpus, parsing

1. Introduction

[1]

Large, searchable, and metadata-rich corpora are basic infrastructure for modern linguistic research. They provide an empirical basis for theoretical investigations, descriptive work such as dictionary and grammar writing, and the development and evaluation of language technology applications. Over the past decade, expectations concerning corpora have partly shifted: alongside carefully sampled reference corpora, continuously expanding, multi-billion-token monitor corpora have become increasingly important. At the same time, scaling up also makes methodological risks more visible, including quality assurance, duplication, access and copyright constraints, and different kinds of bias.

For Hungarian, a sustainable digital presence requires large-scale, curated corpora that are well described with metadata and available to both researchers and developers. A key resource in this area has been the Hungarian Gigaword Corpus (MNSZ2, Oravecz et al., 2014), which was preceded by the first Hungarian National Corpus (MNSZ1, Váradi, 2002). MNSZ2 contains approximately 1.04 billion running words distributed across six domains and includes not only language use from Hungary but also Hungarian texts from neighbouring countries.

The aim of MNSZ3 is to expand this resource by an order of magnitude while preserving the domain structure and balance of MNSZ2. The project is driven by three closely related goals: increasing scale in a controlled way, improving regional coverage – especially for Hungarian used outside Hungary – and establishing a reproducible, maintainable processing workflow for collection, cleaning, annotation, and access.

1.1. International context

Internationally, the label “national corpus” covers several partly different practices: classical balanced reference corpora, continuously updated monitor corpora, and corpora built from multiple subcorpora that cover different time periods and registers. One well-known example of the reference approach is the British National Corpus (BNC), a roughly 100-million-word collection of spoken and written English (BNC Consortium, 2007). Its modern counterpart, BNC2014, was built on a similar scale with a focus on British English of the 2010s (Brezina et al., 2021). At the other end of the spectrum, the German DeReKo contains more than 42 billion units and continues to grow (Kupietz et al., 2018). Czech corpus practice combines a large versioned written corpus (SYN v13) with smaller balanced reference corpora such as SYN2020 (Hnátková et al., 2014; Jelínek et al., 2021). Comparable large national resources also exist for Polish, Spanish, and Russian (Narodowy

[1]This paper is a slightly extended version of Ligeti-Nagy et al. (2025).

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 84–90. Palma, Mallorca. May 11, 2026. ©ELRA Language Resources Association, 2026

Korpus Języka Polskiego (NKJP) / Porowski, S., Bańko, M., et al., 2024; Real Academia Española, 2025; Savchuk et al., 2024).

These examples show that there is no single model for a national corpus. MNSZ3 is positioned between the tradition of balanced reference corpora and the present-day need for substantially larger resources.

1.2. Hungarian background

The direct predecessors of MNSZ3 are the first Hungarian National Corpus (MNSZ1) and the Hungarian Gigaword Corpus (MNSZ2). The common principle behind both was a balanced, domain-based corpus model that aims not at strict statistical representativeness but at a professionally motivated and interpretable balance of text types and sources.

MNSZ1 was designed as a balanced reference corpus of contemporary written Hungarian. Because of practical constraints, spoken language could not yet be included in a comprehensive way, and the available sources were mainly electronically accessible written texts. Even at this stage, however, the corpus design already took the geographical dimension into account and included samples from Hungarian communities in neighbouring countries.

MNSZ2 was developed as a larger and reannotated successor to MNSZ1. Its development was motivated by three main considerations: scale, to support data-driven methods and the study of rare phenomena; quality, through better analysers and finer annotation; and coverage, through resampling and the inclusion of previously underrepresented registers.

1.3. Objectives

The aim of this paper is twofold. First, it presents the motivation and design principles of the MNSZ3 corpus-building programme in the context of earlier Hungarian corpus work. Second, it documents the practical decisions and procedures needed to build a multi-billion-token corpus that remains searchable, reusable, and sustainable.

The project is organised around three goals:

1. **Scaling up with a balanced structure.** The size of MNSZ2 is increased by an order of magnitude while the domain structure and proportions remain controlled.

2. **Strengthening regional coverage.** Hungarian texts from outside Hungary are included in much greater quantities and with richer metadata, especially in those region–domain combinations that were missing or marginal in MNSZ2.

3. **Quality and sustainability.** The project gives greater weight to curated sources, explicitly handles legal and access constraints, and establishes reproducible workflows for collection, cleaning, and versioning.

2. Corpus design principles and structure

This section summarises the main design principles of MNSZ3: the domain system and its definitions, regional and temporal coverage, the metadata model, and the basic requirements of the planned query and access framework.

2.1. Design principles

For national corpora, strict statistical representativeness is not achievable in practice. Hungarian corpus building has therefore traditionally relied on the notion of *balance*. In this approach, the corpus does not claim to mirror reality directly; rather, it models text types and source groups in a proportionate and professionally motivated way, with an explicit domain and partly regional structure (Váradi, 2002; Oravecz et al., 2014). MNSZ3 follows this principle: it preserves the inherited domain structure and proportions of MNSZ2 and scales up within that framework.

A second guiding principle is scalability and reproducibility. The corpus relies on processing schemes that can be applied across large volumes of data and rerun in a controlled manner when the corpus is updated. This requires versioned outputs, stable identifiers, and documented filtering and cleaning steps.

2.2. Domains

The domain system of MNSZ3 follows that of MNSZ2 and consists of six major domains: press, fiction, scientific and popular scientific texts, personal texts, official texts, and transcribed spoken language. The purpose of this system is twofold: it provides broad register coverage and at the same time preserves a modular structure that can be queried at subcorpus level, separately or in combination.

In operationalising the domains, special care is needed for boundary cases, such as the relation between press and opinion writing, or the different genres grouped under official texts. To preserve queryability, domain labels need to be complemented by richer document-type and source metadata.

Table 2 shows the regional distribution of MNSZ2 together with the planned targets for MNSZ3. The

Table 1: Main domains of MNSZ3.

Domain	Definition / typical document types
Press	Edited news and background material: articles, interviews, reports, opinion pieces, lifestyle texts; publisher and source metadata, topical labels.
Fiction	Primarily fictional prose: novels, novellas, short stories; temporal control; translations marked separately.
Scientific / popular scientific	Scientific and popular scientific texts; genre and source metadata; thematic labelling.
Personal	User-generated and personal written communication, such as blogs, forums, and social platforms.
Official	Laws, public and institutional documents, minutes, and related materials; finer sublabelling by document type.
Transcribed spoken language	Transcribed spoken texts, such as conversations, interviews, and broadcasts, with metadata on source and date.

Table 2: Regional distribution of MNSZ2 by domain (million running words) and planned targets for MNSZ3 (million tokens).

Domain	MNSZ2 region (M words) HU	SK	UA	RO	RS	MNSZ2 total	MNSZ3 target	Main focus of expansion
Press	350.5	11.6	0.7	0.6	1.5	364.8	3918	proportional growth; targeted expansion outside Hungary
Fiction	77.0	2.3	0.4	0.8	0.2	80.6	161	moderate growth; better metadata for translations and originals
Scientific	112.0	3.3	0.7	1.6	0.3	117.9	1300	proportional growth; targeted inclusion of institutional sources
Official	98.0	0.2	0.3	0.6	0.03	99.0	1171	targeted expansion of official texts from neighbouring countries; finer document-type labels
Personal	300.3	–	0.4	0.4	0.03	301.1	3011	filling missing Slovakian material; expansion in all regions
Spoken	76.2	–	–	–	–	76.2	836	inclusion of spoken-language material outside Hungary; unified transcription and metadata framework
Total	1013.9	17.3	2.5	3.9	2.0	1039.7	10397	approximately tenfold overall expansion

Note: HU = Hungary, SK = Slovakia, UA = Transcarpathia, RO = Transylvania, RS = Vojvodina. MNSZ2 figures are given in million running words; MNSZ3 targets are in million tokens.

table makes it clear which region–domain combinations were missing or only marginally represented in MNSZ2; their targeted inclusion is one of the priorities of the expansion.

3. Data collection and sources

The expansion of MNSZ3 is not conceived as the creation of an unrestricted web corpus. Instead, it is organised as domain-based data collection that preserves the domain structure and proportions established in MNSZ2 and increases scale within that framework.

Throughout the project, we have aimed to ensure that both source selection and processing follow explicit procedures rather than ad hoc decisions. We fixed temporal coverage domain by domain: for press texts, publication date is the primary reference point; for official texts, the relevant date depends on the document type; for fiction, the date of first publication is used, and for translations, the year of translation. In several domains, 1990 serves as a practical lower boundary, although in fiction we go back to 1945 when necessary to secure enough material under genre constraints.

At document level, we record at least the source, date, and domain, and, where possible, also additional fields such as author, section or topic, and document type. We also anticipated large-scale repetition from the start: identical or near-identical content appears frequently in web and institutional collections, so document-level deduplication is typically combined with boilerplate removal and further domain-specific filtering.

3.1. Press

Press texts were collected through several complementary channels. In web-archive-based harvesting, especially from Common Crawl, we used pre-filtering and content extraction to target article-like pages. We then deduplicated at document level and handled typical cases of near-duplication, such as the same article under different URLs or pages with repeated recommendation blocks. In targeted portal-level harvesting, we aimed at broad coverage of each news site's article inventory and adapted metadata extraction – title, section or topic, author, publication date – to the structure of each

source.

Quality filtering in the press domain consistently excluded index and listing pages, excessively short or incomplete texts, non-Hungarian content, and pages that did not match the genre profile of press articles. The main source groups are the following:

- web-archive material, especially Common Crawl, based on curated press-domain lists;
- targeted harvesting from major Hungarian news sites such as `24.hu`, `hvg.hu`, `nemzetisport.hu`, `blikk.hu`, `vg.hu`, `nlc.hu`, `telex.hu`, `444.hu`, `femina.hu`, and `vezess.hu`;
- inherited or earlier collections already used in MNSZ2;
- press sources from neighbouring countries, for example `hirek.sk`, `gutaonline.sk`, `bulvar.parameter.sk`, `amikassa.sk`, `maszol.ro`, and `hirmondo.ro`.

The current press material collected from Hungarian sources amounts to roughly 2.6 billion tokens from around 100 news domains after deduplication, language filtering, and boilerplate removal. Where topic metadata is incomplete or inconsistent, automatic topic detection is used to assign corpus-wide thematic labels in a consistent way (Osváth and Héja, 2025).

3.2. Official texts

The official domain is built primarily from large structured collections of legal and administrative documents. Extraction and cleaning had to be adapted to several different web and markup environments, including HTML-based sources, structured exports, and TEI-like parallel files. Particular attention was paid to isolating the linguistically relevant text core and excluding headers, footers, repeated navigation elements, attachments, tables, and metadata blocks.

According to the current project records, the official domain reaches close to 0.9 billion words. The main source groups include the following:

- the National Law Repository (approximately 83 million words);
- the Repository of Municipal Decrees (approximately 290 million words);
- anonymised court decisions (approximately 431 million words);
- the Hungarian part of JRC-Acquis (approximately 44 million words);
- the Hungarian side of Europarl (approximately 12 million words);
- the Hungarian subset of DCEP (approximately 35 million words);
- supplementary parliamentary material from Hungary, such as documents and minutes, with internal sublabelling inside the official domain (approximately 79 million words);
- institutional and municipal sources from neighbouring countries.

3.3. Fiction

The fiction domain was planned as a controlled expansion from the start. Here, accessibility, copyright, and quality assurance—especially OCR errors and heterogeneous metadata—set the main limits. For this reason, the collection prioritises prose works with clear dates and verifiable origin, and ambiguous cases such as borderline genres or uncertain translation status are handled in a separate control procedure.

The main sources are the Hungarian Electronic Library (MEK), the Digital Literary Academy (DIA), and the inherited fiction component of MNSZ2. At the current stage, the fiction subcorpus contains 884 documents with a total size of 50,345,981 tokens. Within this material, metadata is maintained both for origin (including texts from outside Hungary and translated works) and for source-side genre labels.

The domain is also one of the most difficult to scale. In practice, the main problems are OCR errors, untidy metadata, encoding issues, and the textual heterogeneity of literary works. Because of these constraints, fiction cannot realistically be expanded at the same rate as the other major domains.

3.4. Scientific texts

In the scientific domain, the most important source group is the REAL repository. These materials were not obtained through direct web harvesting, but were integrated from a parallel project and adapted to the requirements of MNSZ3. The main task here is quality assurance, because a large share of the material is OCR-based and therefore contains character substitutions, line-break errors, hyphenation artefacts, and intrusive headers, footers, and page numbers.

Cleaning is carried out in several steps: character encoding and basic typography are normalised, repeated non-content elements are removed, rule-based corrections are applied to common OCR patterns, and the corrected texts are added to the corpus after quality control. At present, the scientific domain of MNSZ3 contains approximately 1.2 billion words.

3.5. Personal texts

Since MNSZ2, access conditions to social-media platforms have changed substantially. This affects one of the key source types for the personal domain: large platforms have become much harder to harvest because of stricter usage terms and data-handling practices. For this reason, the personal domain relies mainly on publicly accessible and stably citable textual sources where document-level processing and minimal metadata recording are feasible.

In practice, blogs are the main source group. A central source is `blog.hu`, harvested in a targeted way and then cleaned by removing navigation, comment, and recommendation blocks, followed by language filtering and deduplication. Where possible, the corpus distinguishes between post-level and comment-level material. In addition to blogs, we also include Reddit, which is still technically accessible for this purpose. The current Reddit-based personal material is on the order of 10 million words.

3.6. Transcribed spoken language

The spoken-language domain is currently being expanded mainly through podcast-like recordings with longer stretches of continuous speech. The main source at present is the podcast collection of the National Széchényi Library.

For transcription, we use an uploader and controller tool attached to the BEAST2 system (Kádár et al., 2023). The program uploads the audio and stores the machine transcripts in a documented directory structure, from which they are reintegrated into the corpus-building pipeline. Automatic post-correction is then carried out with a dedicated correction component designed to address typical transcription problems such as mishearing, punctuation, sentence boundaries, and unstable spelling of proper names.

3.7. Hungarian texts from neighbouring countries

Regional extension started from web-archive-based domain lists that contained Hungarian-language content. From these lists, we selected domains associated with neighbouring countries and assigned them manually to the MNSZ3 domains. Because the sources are highly heterogeneous, project-specific harvesting and extraction scripts were developed. Text and metadata were stored separately, then filtered for language, deduplicated, and cleaned of boilerplate content.

In the press domain, the material currently collected from neighbouring countries contains 232,323,495 tokens and 1,206,612 documents. Most of this material comes from Romania and Slovakia, with smaller but still relevant collections from Austria, Croatia, Serbia, Slovenia, and Ukraine.

In the official domain, the corresponding material currently contains 36,094,964 tokens and 114,971 documents, again with the largest share coming from Romania and additional contributions from Slovakia, Ukraine, and Serbia.

In the personal domain, collection is still in progress. The available material comes mainly from open web sources such as blogs, personal homepages, service-oriented self-presentations, political personal pages, and some forum-like sites. Because the source base remains heterogeneous and uneven across countries, no final aggregate figures are given for this component at the current stage.

4. Linguistic annotation

One of the main commitments of MNSZ3 is that it will not only be large and balanced, but also linguistically analysed. In addition to tokenisation and lemmatisation, the corpus is planned to provide morphosyntactic information, dependency parsing, named entity recognition, and keyword extraction. The goal of annotation is twofold: to improve corpus usability for search and linguistic investigation, and to establish a technological basis for further components such as terminology extraction and domain-specific normalisation.

For large-scale processing, we compared two widely used Hungarian language-processing pipelines: e-magyar (Váradi et al., 2017; Indig et al., 2019) and HuSpaCy (Orosz et al., 2022, 2023). The two systems overlap in several functions, but differ in architecture and strengths. E-magyar is a modular, research-oriented framework whose most distinctive component is the rich emMorph-based morphological analysis (Novák et al., 2016). HuSpaCy, by contrast, is built on spaCy (Honnibal et al., 2020), provides a unified UD-oriented pipeline, and is well suited to efficient large-scale processing.

Because annotation in MNSZ3 means running the analysis on billions of words, individual component quality is not the only consideration. Output consistency, reproducibility, error handling, and stability across domains are equally important. To support the comparison, we built a unified evaluation framework (*Launcher*) that runs both analysers on the same input, applies minimal normalisation where necessary, compensates for tokenisation shifts heuristically, and returns differences in a form that supports both aggregation and manual inspection. This part of the work is based on Skrabák and Ligeti-Nagy (2025).

4.1. Test material and main findings

The comparison was carried out on a parliamentary-record sample. This material is both linguistically varied and formally heterogeneous, which makes it a suitable stress test for tokenisation, morphology, and syntax. In the sample, HuSpaCy identified 46,607 tokens and e-magyar 46,452 tokens, with the difference largely attributable to recurring tokenisation patterns.

The results can be summarised as follows.

- **Tokenisation.** HuSpaCy performed better overall. Weighted by frequency of occurrence, it gave the better solution in 89.9% of all differing cases and in 93.9% of the unambiguous ones.
- **Morphology and lemmatisation.** E-magyar provided the more reliable output. In the HuSpaCy–emMorph integration, 4,987 out of 46,635 tokens returned `None` at the morphological level. Even after corrections to the integration, substantial differences remained between the outputs of the two systems.
- **Part-of-speech tagging.** Many differences were driven by tokenisation and punctuation handling rather than purely by tag assignment.
- **Dependency parsing.** HuSpaCy produced the more consistent and reusable output. In e-magyar, some sentences lacked a ROOT relation, which is a serious problem for downstream processing.
- **Named entity recognition.** HuSpaCy tended to identify more named-entity tokens, while e-magyar was more conservative but more precise in type assignment on the manually checked sample.

4.2. Annotation strategy

The comparison suggests that the two pipelines should not be seen as a simple "better versus worse" contrast. Rather, they have complementary strengths. HuSpaCy is stronger in tokenisation, dependency parsing, and named entity recognition; e-magyar provides better morphology and lemmatisation.

The annotation strategy of MNSZ3 therefore combines them. HuSpaCy provides the unified and efficient backbone for large-scale UD-oriented processing, including tokenisation, POS/UD morphology, dependency parsing, and named entity recognition. At the same time, the emMorph-based output of e-magyar is preserved as an explicit additional layer for morphology and lemmatisation. In parallel, the HuSpaCy–emMorph integration remains an independent line of development, with the aim of bringing the morphological and lemmatisation behaviour of the HuSpaCy-based workflow closer to that of e-magyar.

5. Conclusion

MNSZ3 is intended to provide a qualitative step forward for the digital presence and research infrastructure of Hungarian. A multi-billion-token corpus that is curated, consistently described with metadata, and linguistically analysed will support both traditional corpus-based linguistic work and present-day language technology. It will allow more reliable investigation of rare phenomena, finer tracking of genre and diachronic differences, and broader inclusion of registers that were previously only marginally available.

At the same time, the project has direct practical value. A more balanced domain structure, stronger representation of Hungarian used outside Hungary, and the extension of spoken-language material all improve the range of linguistic variation available for study. Detailed metadata and a reproducible processing workflow improve transparency and reusability, and lower the practical threshold for corpus use among researchers, developers, and educators.

6. Bibliographical References

BNC Consortium. 2007. British National Corpus, XML Edition. Available at `http://www.natcorp.ox.ac.uk/`,.

V. Brezina, A. Hawtin, and T. McEnery. 2021. The written British National Corpus 2014 – design and comparability. *Text & Talk*, 41(5-6):595–615.

Milena Hnátková, Michal Křen, Pavel Procházka, and Hana Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Péter Kundráth, and Noémi Vadász. 2019. emtsv – Egy formátum mind felett. In *Magyar Számítógépes Nyelvészeti Konferencia*, Szeged. In Hungarian.

Tomáš Jelínek, Jan Křivan, Vladimír Petkevič, Hana Skoumalová, and Jitka Šindlerová. 2021. SYN2020: A new corpus of Czech with an innovated annotation. In *Text, Speech, and Dialogue*, Lecture Notes in Computer Science, pages 48–59. Springer.

Máté Soma Kádár, Gergely Dobsinszki, Katalin Mády, and Péter Mihajlik. 2023. "Feeding the BEAST" – A BEA Speech Transcriber továbbfejlesztése és integrálása neurális nyelvmodellel. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia*, volume 19, pages 135–143.

Marc Kupietz, Harald Lüngen, Paweł Kamocki, and Andreas Witt. 2018. The German reference corpus DeReKo: New developments – new opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4353–4360, Miyazaki, Japan. European Language Resources Association (ELRA).

Noémi Ligeti-Nagy, Enikő Héja, Ágnes Bánfi, Flóra Földesi, Mariann Lengyel, Bence Sárossy, Boglárka Skrabák, Tamás Váradi, and Gábor Prószéky. 2025. Expanding the Hungarian Gigaword Corpus. In *CLARIN Annual Conference Proceedings 2025*, pages 188–192, Vienna. CLARIN ERIC.

Narodowy Korpus Języka Polskiego (NKJP) / Porowski, S., Bańko, M., et al. 2024. Narodowy korpus języka polskiego. Dostępne online na https://nkjp.pl/.

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1719–1723. European Language Resources Association (ELRA).

György Orosz, Gergő Szabó, Péter Berkecz, Zsolt Szántó, and Richárd Farkas. 2023. Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In *Text, Speech, and Dialogue*, pages 58–69, Cham. Springer Nature Switzerland.

György Orosz, Zsolt Szántó, Péter Berkecz, Gergő Szabó, and Richárd Farkas. 2022. HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 59–73.

Mátyás Osváth and Enikő Héja. 2025. Internetes hírek automatikus osztályozása. In *Magyar Számítógépes Nyelvészeti Konferencia (21.)*, volume 21, pages 29–39, Szeged. Szegedi Tudományegyetem TTIK, Informatikai Intézet. Konferenciaközlemény. Elérhető: http://acta.bibl.u-szeged.hu/id/eprint/88770.

Real Academia Española. 2025. CORPES XXI: Corpus del Español del Siglo XXI. Accessed: 22/01/2026.

S. O. Savchuk, T. Arkhangelskiy, A. A. Bonch-Osmolovskaya, O. V. Donina, Yu. N. Kuznetsova, O. N. Lyashevskaya, B. V. Orekhov, and M. V. Podryadchikova. 2024. Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, (2):7–34.

Boglárka Skrabák and Noémi Ligeti-Nagy. 2025. A huspacy és e-magyar elemzőláncok teljesítményének átfogó összehasonlítása országgyűlési szövegeken: a tokenizálástól a függőségi elemzésig. In *Magyar Számítógépes Nyelvészeti Konferencia*, volume 21, pages 171–184, Szeged. Szegedi Tudományegyetem TTIK, Informatikai Intézet.

Tamás Váradi. 2002. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 385–389.

Tamás Váradi, Eszter Simon, Bálint Sass, Mátyás Gerőcs, Iván Mittelholtz, Attila Novák, Balázs Indig, Gábor Prószéky, and Veronika Vincze. 2017. Az e-magyar digitális nyelvfeldolgozó rendszer. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, pages 49–60, Szeged. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

CoRoLa version 2.0: Corpus Enrichment and a New Annotation Level

Elena Irimia, Verginica Barbu Mititelu, Radu Ion, Vasile Păiș, Maria Mitrofan, Dan Tufiș

Research Institute for Artificial Intelligence, Romanian Academy, Romania

{elena,vergi,radu,vasile,maria,tufis}@racai.ro

Abstract

The paper gives an overview of the recent developments in the enrichment of the reference Corpus of Contemporary Romanian (CoRoLa), within on-going international projects. Statistics of the newly acquired data, work methodology and work towards inclusion of a new annotation layer, the syntactic one, are detailed. We briefly present RODNA, an updated Romanian text processor with state-of-the-art performance on POS tagging, lemmatization and dependency parsing that will be used to populate the syntactic layer of CoRoLa.

Keywords: Romanian, corpus, NLP pipeline, RODNA

1. Introduction

Language resources in the form of large language corpora are still valuable assets provided that they are carefully curated and offer access to texts, metadata and annotation information inaccessible to the Large Language Models (LLMs) already widely available. In the current context, when Artificial Intelligence generated texts have become ubiquitous, corpora that guarantee the originality of their content and offer access to their content have high chances to become valuable repositories of the natural languages characteristics.

We present below the Reference Corpus of Contemporary Romanian (CoRoLa) (Barbu Mititelu et al., 2018), which was launched in 2017 and now is being under further quantitative and qualitative enrichment: new texts are collected, the design of the metadata is adjusted to reflect the newly added types of texts, and a new annotation level is added to the whole corpus. This will represent version 2.0 of CoRoLa.

Section 2 below describes the main characteristics of CoRoLa version 1.0, at the moment of its release in 2017, while the steps taken towards its version 2.0 are described in Section 3. This details the contexts in which new texts are collected and the methodology adopted for this mainly automatic collection. The major step taken in the corpus development is its syntactic parsing, using the dependency grammar, and this is the focus of Section 4, before concluding the paper.

2. CoRoLa 1.0

The development of a reference corpus to reflect the contemporary Romanian language was a priority project of the Romanian Academy, carried out by two of its institutes (Research Institute for Artificial Intelligence from Bucharest and the Institute for Computer Science from Iași). However, this was a national wide endeavour, as, on the one hand, collecting texts meant (and still means nowadays, given the lack of adaptation of the legislation to the evolution of technology and research) contacting publishers, media representatives and other decision makers in order to get access to the data. On the other hand, for their processing, human resources were needed, thus universities around the country were also contacted and they agreed to have their students involved in metadata creation and data cleaning.

Besides national bodies, the project also had an international component: it was due to the DRuKoLA project[1] (funded by the Alexander von Humboldt Foundation) that CoRoLa benefited from a reliable infrastructure to index its content and offer query-based access to it, i.e. the KorAP Corpus Query Platform (Diewald et al., 2016).

2.1. Design of CoRoLa

At the moment of its development, the corpus was meant to answer needs of several communities: linguists, for which the corpus is a source of various language phenomena attestation, offering a glimpse of their frequency, of the characteristics of various language styles, domains, etc.; language engineers, for which the corpus was a source of word embeddings (see below for those extracted from CoRoLa); the public, who can find here original, natural uses of various words, alongside their collocations.

As a work methodology, we gathered texts from various sources (from books to online newspapers, from textbooks to poetry, etc.) in various formats (PDF, DOC(X), MP3, WAV files), and both automatically and manually. Most of the collected data could be processed, cleaned and added to the corpus, while a small part of it proved unusable, as text could not be extracted from some files.

[1] https://www.ids-mannheim.de/digspra/pb-s1/projekte/drukola/

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 91–97. Palma, Mallorca. May 11, 2026.

2.2. Statistics of CoRoLa v. 1.0

The first version of CoRoLa comprised, in its written component, 1,257,752,812 tokens, unevenly distributed across multiple language styles (legal, administrative, scientific, journalistic, imaginative, memoirs and blog posts), four domains (arts and culture, nature, society, science) and 71 subdomains. The corpus was processed with the in-house tool TTL (Ion, 2007) for sentence splitting, tokenization, morpho-syntactic annotation and lemmatization, achieving an accuracy of approximately 97.5% (Tufiș et al., 2008). Documents were indexed in a local instance of the KorAP[2] corpus query and analysis platform (Diewald et al, 2019). Word embeddings[3] (Păiș and Tufiș, 2018) and frequency lists[4] computed on CoRoLa were also made public.

Moreover, the corpus has an oral component containing about 300 hours of recordings with transcriptions, which is also available for querying[5].

If at the moment of its launching, in 2017, a corpus of 1.2 billion tokens, entirely I(ntellectual)P(roperty)R(ights)-cleared, manually classified and validated, was considered a remarkable achievement, today this size is no longer impressive. At the same time, compliance requirements with IPR regulations remain stringent, and therefore the difficulties in collecting relevant data are still considerable. However, efforts to obtain usage rights for new data have proven successful, as shown below.

In what follows, we present new high-quality datasets that have been included or will be included in version 2.0 of CoRoLa. The data is both extensive and highly diverse, improving the initial distribution across linguistic styles and domains. In addition, new corpus processing tools, together with the inclusion of additional language varieties, enable broader investigations and extended applications.

2.3. Uses of CoRoLa

Over time, CoRoLa has supported:

- linguistic research (offering quality empirical data for theoretical studies: Ștefănescu, 2019; Ștefănescu et al., 2020; Giurgea, 2024; Bîlbîie, 2025; Vasileanu and Niculescu-Gorpin; 2025),
- the development of comparable corpora within projects such as: DruKoLa, which devised a pilot German-Romanian comparable corpus for contrastive linguistic analysis (Kupietz et al., 2019), CURLICAT[6] (Váradi et al., 2022), in which 3,042 documents from CoRoLa were IPR-cleared for further distribution by getting back to text providers with new agreement proposals
- Natural Language Processing tasks, such as Named Entity Recognition: CoRoLa-based embeddings were used to enhance a general named entity recognizer for Romanian implemented using Conditional Random Fields (CRF), based on the Stanford NER (Finkel et al., 2005) software package.

3. First Steps into CoRoLa 2.0

Recent international projects offered the perfect opportunity to resume corpus expansion, with particular care to exclude AI-generated content by restricting web crawling to texts published earlier than 2023.

3.1. MARCELL

In the project *Multilingual Resources for CEF.AT in the legal domain*[7] (MARCELL), funded by the Connecting Europe Facility, a large quantity of Romanian legal texts (163,000 files, 4,434,000 sentences, 412,000,000 tokens, which represent the body of national legislation ranging from 1881 to 2021) was collected and, given that such texts are not under any usage restrictions, they can be easily added to CoRoLa. All the texts were obtained via crawling from the public Romanian legislative portal[8]. We have not distinguished between "in force" and "out of force" laws because it is difficult to do this automatically and there is no external resource to use to distinguish between them. The texts were extracted from the original HTML format and converted into TXT files, metadata was automatically created and the texts underwent automatic processing and morpho-syntactic annotation.

3.2. LLMs4EU

Large Language Models for the European Union[9] (LLMs4EU) is a European Commission-funded project bringing together 66 European partners, including companies and research institutions, under the coordination of ALT-EDIC[10]. The project seeks to ensure the availability of LLMs and the necessary tools for their deployment across all EU languages by building on existing European programs and expertise and promoting open-data access. Within this framework, our team contributes by collecting new corpus data and providing fine-tuning and evaluation datasets, including resources derived from CoRoLa.

A set of 30,584 legal documents were extracted from the same Romanian legislative portal. The collected texts are from the period 2022-2025

[2] https://korap.racai.ro/
[3] https://corolaws.racai.ro/word_embeddings/
[4] https://zenodo.org/records/7091535
[5] http://89.38.230.23/corola_sound_search/index.php
[6] https://curlicat-project.eu/
[7] https://marcell-project.eu/
[8] https://www.just.ro/
[9] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-details/43152860/101198470
[10] https://language-data-space.ec.europa.eu/related-initiatives/alt-edic_en

(therefore not overlapping the data from MARCELL) and contain 2,383,809 sentences comprising 77,572,697 tokens. These IPR-cleared texts were annotated using UDPipe (Straka et al, 2016), and indexed in CoRoLa under a dedicated LLMs4EU sub-corpus label.

Also as part of the LLMs4EU project, we have extracted a corpus of Romanian language doctoral theses from the national Integrated Educational Registry platform (REI)[11], spanning thirteen broad scientific domains. The corpus comprises a total of 12,523 doctoral theses and over 770 million words (see Table 1), reflecting a cross-section of Romanian academic production. The largest contributions come from medicine and interdisciplinary fields, followed by engineering and humanities (see Table 1 for the exact numbers of documents and words as per each domain represented in the data). Smaller yet significant contributions are represented by arts, economics, theology, natural sciences, social sciences, life sciences, security and defense, law, and computer science. The total number of collected documents and that of words are also presented in Table 1.

At present, these theses are published under the Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International license (CC BY-NC-ND 4.0), which permits redistribution with attribution but prohibits commercial use and the creation or distribution of derivative works. Since NLP (Natural Language Processing) operations, including tokenization, fine-tuning, and the construction of training datasets for large language models, constitute derivative transformations under the terms of this license, we are currently in the process of contacting the host institution in order to obtain explicit permission to apply such derivative operations on these materials and include them in CoRoLa.

Domain	# docs	# words
medicine	3,654	176,455,463
interdisciplinary	2,507	174,166,039
engineering	1,724	83,558,112
humanities	664	66,858,211
arts	752	50,330,127
economics	683	47,056,316
theology	338	38,794,193
natural_sciences	615	30,854,763
social_sciences	364	24,190,568
life_sciences	402	21,070,967
security_defense	267	19,997,678
law	166	19,453,414
computer_science	387	17,763,040
TOTAL	**12,523**	**770,548,891**

Table 1: PhD thesis statistics.

Within the same project, we have identified new sources of quality data and have signed protocols for data collection, storing and processing. One such protocol was signed with "G. Călinescu" Institute for Literary History and Theory[12] and 51 IPR cleared files (containing 17, 525,736 words) have been received that are currently preprocessed in order to be included in CoRoLa. The documents have academic content from the philology and literary theory field.

Another batch of IPR-cleared files come from publishing houses and contain 2,981,620 words.

Altogether, within LLMs4EU texts comprising 868,628,944 words have been collected so far. They are in different processing steps, but they are all on their way into CoRoLa.

3.3. ADAMo

In the bilateral Romanian-Moldovan project *Automatic Detection of AI-Generated Texts from Moldova and Romania*[13] (ADAMo) our aim is to develop a classifier capable of identifying Artificial Intelligence (AI)-generated texts with characteristics from any of the two varieties of the Romanian language. Even though both countries have the same national language, Romanian, there are important differences between the variants spoken therein. For the first time, these specific features will be automatically identified at different language levels within this project, including the syntactic one, which is an annotation layer that will be added to CoRoLa within the ADAMo project.

At the moment, 12 million tokens (from the targeted 15 million) of high-quality, IPR-cleared (written, as well as oral) texts representative of the Moldovan variety of Romanian have already been collected in ADAMo and they will be added to CoRoLa, with adequate metadata, consistent with the ones for the texts already in the corpus.

Both the newly collected corpus and comparative texts extracted from CoRoLa will be used for developing the classifiers able to distinguish, on the one hand, between texts belonging to the two different language varieties, and, on the other, between original and AI-generated texts.

3.4. DeepNewDef

In the context of the project *Defending against deep fake news with large language and image models*[14] (DeepNewDef), we are building tools for

[11] https://rei.gov.ro/
[12] https://www.inst-calinescu.ro/
[13] https://www.racai.ro/p/adamo/index.html
[14] https://www.racai.ro/p/deepnewsdef/index.html

detecting fake news content (both text and images), considering the specifics from Romania and the Republic of Moldova.

In this context, the CoRoLa corpus will be used as a source of human-written text for training fake text detection models. Furthermore, as new content will be gathered throughout the project, original and IPR-cleared parts of it will be indexed and made available in CoRoLa.

4. Processing CoRoLa 2.0

The corpus will be reprocessed with RODNA[15] (**RO**manian **D**eep **N**eural networks **A**rchitectures, Ion, 2022), an actively developed, Romanian-specific NLP pipeline that performs sentence splitting, Romanian-aware tokenization, fine-grained Part-of-Speech (POS) tagging, lemmatization, and dependency parsing with Universal Dependencies (de Marneffe et al., 2021) dependency relations.

RODNA implements a Romanian-aware tokenizer that uses rules to correctly and consistently split dash-affixed clitics (e.g. "să-ți", "purtându-mi-l"), to recognize different types of numbers (real, percentages, integers grouped by three digits), time and dates (e.g. "12:15", "25/12/2025"), and abbreviations (e.g. "F.I.F.A.").

RODNA's lemmatizer uses a large Romanian lexicon[16] (more than 1.1M wordforms) in which each wordform is listed with its fine-grained POS tag (called a Morpho-Syntactic Descriptor or MSD[17]) and its lemma for that POS tag. Lemmatization is performed after POS tagging, so that we can get a set of lemmas for the pair wordform and MSD. If this set has more than one element, the most frequent lemma is used. When the word is not in the lexicon, it is assumed to be a content word (i.e. noun, verb, adjective or adverb) and lemmatization is obtained via the Romanian Paradigmatic Morphology (Irimia, 2009), which RODNA also implements.

For sentence splitting and POS tagging, RODNA uses a Romanian BERT model (Dumitrescu et al., 2020) to provide input embeddings for specialized neural networks heads that perform classification:

- Sentence splitting is done with a bidirectional LSTM (Long Short-Term Memory) network, that classifies a token as bearing the "end of sentence" mark or not.
- POS tagging is performed using the "tiered tagging" methodology (Tufiș and Dragomirescu, 2004): each token is first classified using a coarse-grained POS tagset, and then, the MSD is extracted from the Romanian lexicon, with a deterministic mapping from the pair (wordform, coarse-grained tag) to the MSD. If the word is not in the lexicon, the most probable MSD is assigned with a neural network that learns to map character embeddings of lexicon words to their possible MSDs.

RODNA's dependency parsing is realized in two steps:

1. construct the unlabeled dependency tree of the input sentence and
2. label each root-to-leaf path in this tree with Romanian UD dependency labels.

Step 1 learns a probability distribution of possible heads of the current token, as relative offsets (in number of tokens) to the left/right of the current token. Each token input is the BERT embedding for it, and we stack a bidirectional LSTM on top to learn a probability distribution of possible heads over a window of $\pm k$ tokens around the target token. Finally, using the Chu-Liu-Edmonds' algorithm for finding the maximum spanning tree in a directed graph (Chu and Liu, 1965; Edmonds, 1967) we obtain the unlabeled dependency tree of the input sentence.

Step 2 takes each root-to-leaf path in the unlabelled dependency tree and considers it an ordered sequence of BERT embeddings corresponding to tokens in the nodes. It learns to label each root-to-child edge by employing a unidirectional GRU neural network taking as input BERT embeddings and outputting a probability distribution over dependency labels.

RODNA has been trained on the "train" split of the Romanian Reference Treebank (RRT, Barbu Mititelu et al., 2016) and evaluated on the "test" split of this corpus, because the "dev" split was used to determine the best model to save, depending on the performance measure of the task on this split.

We compare RODNA to state-of-the art, multilingual text processing tools such as Stanza (Peng et al., 2020) and Trankit (Nguyen et al., 2021). Both text processors have been trained on the current version of the RRT "train" split and both of them use the "dev" split to choose and save their best models during training iterations.

Stanza also allows the integration of BERT embeddings; accordingly, it was trained using the same Romanian BERT model employed for RODNA. In contrast, Trankit relies exclusively on flavours of the XLM-RoBERTa model and does not support training with alternative BERT architectures.

Comparison is done by running all text processors on the raw text of the "test" split and letting each processor do sentence splitting, tokenization, POS tagging, lemmatization and dependency parsing. In order to compare them fairly, we align their outputs at sentence and token level with the gold standard, and compute accuracy for POS tagging and lemmatization, and Unlabelled

[15] https://github.com/racai-ai/Rodna

[16] https://github.com/racai-ai/Rodna/blob/master/data/resources/tbl.wordform.ro

[17] https://nl.ijs.si/ME/V6/msd/html/msd-ro.html

Attachment Scores (UAS) and Labelled Attachment Scores (LAS) for dependency parsing. In order to back up the claim that one processor is better than other processor, we use the McNemar's paired test to verify the null hypothesis that $n_{10} \approx n_{01}$, that is, the number of times the first text processor is correct when the second is not is about the same as the number of times the second text processor is correct when the first is not. n_{11} gives us the number of times both text processors are correct while n_{00} is the number of times neither is correct.

	n11	n00	n10	n01	Rodna	Stanza	Null
CG	14352	168	102	88	98.26%	98.16%	No
FG	14274	214	122	100	97.86%	97.71%	No
Lm	14337	88	205	80	98.85%	98%	Yes
US	12944	683	536	547	91.63%	91.71%	No
LS	12115	1177	684	734	87%	87.34%	No

Table 2: RODNA vs. Stanza on the "test" split of RRT

Table 2 shows the coarse-grained POS tagging accuracy (CG, with UD UPOS tags), the fine-grained POS tagging accuracy (FG, with MSDs), lemmatization accuracy (Lm), the UAS percentage (US) and the LAS percentage (LS). Underlined values are higher, but the Null column shows if the null hypothesis can be rejected or not, that is, if RODNA is significantly better than Stanza or the other way around, at a p-value of 0.05. Currently Rodna significantly outscores Stanza only when doing lemmatization because it uses the Romanian Paradigmatic Morphology when lemmatizing unknown words.

	n11	n00	n10	n01	Rodna	Trankit	Null
CG	14884	155	100	109	98.26%	98.32%	No
FG	14765	209	161	113	97.88%	97.53%	Yes
Lm	13997	117	1078	56	98.86%	92.16%	Yes
US	13138	897	631	582	90.3%	89.97%	No
LS	12310	1403	771	764	85.78%	85.74%	No

Table 3: RODNA vs. Trankit on the "test" split of RRT

With respect to Trankit, RODNA is significantly better when doing fine-grained POS tagging and especially when doing lemmatization: Trankit lemmatizer is subpar, when also tested with pre-trained models that it automatically downloads.

RODNA outputs files in the CoNLL-U format[18] which will be indexed in KoRAP, in a specific RODNA foundry, following KorAP's multi-layer indexing model (Diewald et al., 2016). This ensures that the lemmatisation, POS tagging and dependency parsing information, although accessible on different annotation layers, remains structurally aligned and can be queried within the same positional index.

By the end of 2026, we foresee the release of CoRoLa 2.0, with the following approximate counts, including the Moldovan variety of Romanian:

Domain	Style	No. of tokens
Law	Law	491,000,000
Science	Academic	790,000,000
-	Imaginative/Memoirs	3,000,000
-	Journalistic	13,000,000
TOTAL		1,297,000,000

Table 4: Projected counts for the CoRoLa 2.0 release.

5.Conclusions

The reference corpus of contemporary Romanian, CoRoLa, is an actively developed resource. All activities carried out towards its enrichment consider data quality and the possibility to use the processed texts in further resources, applications and downstream tasks development. CoRoLa is gaining language varieties representation (both at the written and the oral level), new types of texts (a particular type of academic one, namely doctoral theses) and is also harnessed in applications development specific to current days.

6. Acknowledgements

This work received support from: (i) a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV, (ii) the "Large Language Models for the European Union (LLMs4EU)", project no. 101198470, call DIGITAL-2024-AI-B-06-LANGUAGE, funded by the European Union, (iii) a grant of the Ministry of Research, Innovation and Digitalization - UEFISCDI, project number PN-IV-P8-8.2-EUD-2025-0061, within PNCDI IV, (iv) NATO Science for Peace and Security Programme under grant id. G8648 (v) a grant of the Ministry of Education and Research, CCCDI - UEFISCDI, project number PN-IV-P8-8.2-NATO-SPS-2025-0005, within PNCDI IV. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

7. Bibliographical References

Barbu Mititelu, V., Ion, R., Simionescu, R., Irimia, E., and Perez, C. A. (2016). The Romanian treebank annotated according to universal

[18] https://universaldependencies.org/format.html

dependencies. In *Proceedings of the tenth international conference on natural language processing (hrtal2016)*.

Barbu Mititelu, V. (2018). Modern syntactic analysis of Romanian. In O. Ichim, L. Botoșineanu, D. Butnaru, M.-R. Clim, O. Ichim, & V. Olariu (Eds.), *Clasic și modern în cercetarea filologică românească actuală*, pp. 67–78. Iași: Publishing House of "Alexandru Ioan Cuza" University.

Barbu Mititelu, V., Tufiș, D., and Irimia, E. (2018). The reference corpus of the contemporary Romanian language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association. https://aclanthology.org/L18-1189/.

Bîlbîie, G. (2025) 'Multiple wh-questions in Romanian: A corpus-based approach', *AND CORPORA*, p. 33. https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Konferences/2025/GGC-10-ba.pdf#page=34

Chu, Y.J. (1965). On the shortest arborescence of a directed graph. *Scientia Sinica*, *14*, pp.1396-1400.

de Marneffe, M.C., Manning, C., Nivre, J. and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics* 47(2): 255–308.

Diewald, Nils, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt (2016). "KorAP Architecture—Diving in the Deep Sea of Corpus Data." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3586-3591.

Diewald, N., Barbu Mititelu, V., and Kupietz, M. (2019). The KorAP user interface. Accessing CoRoLa via KorAP. *Revue Roumaine de Linguistique. 64*(3), 265-277

Dumitrescu, S., Avram, A. M., & Pyysalo, S. (2020). The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4324-4328).

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the national Bureau of Standards B*, *71*(4), pp.233-240.

Finkel, J.R., Grenager, T. and Manning, C.D., (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pp. 363-370.

Giurgea, I. (2024). Romanian double definites: The view from demonstratives. *Lingua*, *307*, p.103728.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD Thesis, Romanian Academy.

Ion, R. (Ed.) *Evaluating and User-Testing Rodna, A New Romanian Text Processing Pipeline*; Research report; Romanian Academy: Bucharest, Romania, 2022.

Irimia, E. (2009). ROG – A Paradigmatic Morphological Generator for Romanian. In *Vetulani, Z., Uszkoreit, H. (eds) Human Language Technology. Challenges of the Information Society. LTC 2007*. Lecture Notes in Computer Science, vol 5603. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04235-5_7

Kupietz, M., Cosma, R. and Witt, A. (2019). The Drukola Project. *Revue Roumaine de Linguistique. On design, creation and use of of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*, *64*(3), pp. 255-263.

Van Nguyen, M., Lai, V. D., Veyseh, A. P. B., & Nguyen, T. H. (2021). Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: System demonstrations,* pp. 80-90.

Păiș, V. and Tufiș, D. (2018). Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy, series A*, *19*(2), pp.403-409.

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 Association for Computational Linguistics (ACL) System Demonstrations. 2020.

Ştefănescu, A. (2019). The Use of Altminteri 'Otherwise'in Romanian: From Adverb to Textual Connector. In *Fuzzy Boundaries in Discourse Studies: Theoretical, Methodological, and Lexico-Grammatical Fuzziness* (pp. 287-313). Cham: Springer International Publishing.

Ștefănescu, A., Postolea, S. and Barbu Mititelu, V. (2020). The Romanian discourse markers *de altfel* and *de altminteri*: Patterns of use and core functions, *RRL*, 65(3), pp. 307–322.

Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4290-4297). European Language Resource Association (ELRA).

Tufiș, D., Ion, R., Ceauşu, A., Ştefănescu D. (2008). RACAI's Linguistic Web Services. In *Nicoletta Calzolari et al. (Eds.) Proceedings of the 6th LREC, Marrakech, Morocco*, European Language Resources Association (ELRA).

Tufiș, D. & Dragomirescu, L. (2004, May). Tiered tagging revisited. In *Proceedings of the 4th LREC Conference. Lisbon, Portugal* (pp. 39-42). Language Resources Association (ELRA).

Váradi, T., Nyéki, B., Koeva, S., Tadić, M., Štefanec, V., Ogrodniczuk, M., Nitoń, B., Pęzik, P., Mititelu, V.B., Irimia, E. and Mitrofan, M. (2022). Introducing the CURLICAT corpora: seven-language domain specific annotated corpora from curated sources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 100-108). European Language Resources Association (ELRA).

Vasileanu, M. and Niculescu-Gorpin, A.-G. (2025) ‘Romanian libfixes in the making’, in Arndt-Lappe, S. and Filatkina, N. (eds.) *Dynamics at the lexicon-syntax interface: Creativity and routine in word-formation and multi-word expressions*. Berlin: De Gruyter, pp. 241–265.

The German Medical Text Corpus: Early 2026 Update

Justin Hofenbitzer[1], **Christina Lohr**[2], **Frank Meineke**[2], **Markus Loeffler**[2], **Martin Boeker**[1]
[1]TUM University Hospital, School of Medicine and Health, Chair of Medical Informatics, Institute for AI and Informatics in Medicine, Technical University of Munich, Germany,
[2]Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Germany

justin.hofenbitzer@tum.de, martin.boeker@tum.de

Abstract

Clinical text resources are a central component for the study of medical language, as well as the training and evaluation of large language models, chatbots, and artificial intelligence systems supporting clinical routines. With the German Medical Text Corpus (GeMTeX), we are currently working on the largest shareable clinical document dataset in German. The multi-centric project ensures diversity across different university hospitals, clinical domains, and text sorts. After a thorough de-identification process, the clinical texts are semantically annotated using Snomed CT, a language-independent, standardized medical ontology. While the corpus is still under active development, it is accessible upon request under controlled access conditions. As of February 2026, GeMTeX comprises more than 15k documents and 20M tokens. We refer researchers interested in the resource to visit https://kiinformatik.mri.tum.de/en/gemtex or reach out to us via gemtex.mi@mh.tum.de.

1. Introduction

The German Medical Text Corpus (GeMTeX) project is a three-year (2023–2026) initiative to establish a multi-site corpus of written German clinical routine text enriched with an ontology-grounded semantic layer (Meineke et al., 2023; Faller et al., 2025). Embedded in the German *Medical Informatics Intitative* (MII) (Semler et al., 2018), the project is organized as a consortium of 18 partners, including six university hospitals contributing documents and annotations, i.e., *TUM Klinikum*, *Universitätsklinikum Leipzig*, *Universitätsklinikum Erlangen*, *Charité Berlin*, *Universitätsklinikum Carl Gustav Carus Dresden*, and *Universitätsklinikum Essen*. GeMTeX addresses two bottlenecks: The scarcity of large German clinical corpora and the limited availability of semantic layers supporting evaluation beyond surface-form matching (Névéol et al., 2018; Hahn, 2025). GeMTeX therefore aims to (i) create a standardized German clinical text corpus, (ii) enable cross-site analyses and reproducible benchmarking, and (iii) provide a Snomed CT[1]-grounded semantic annotation layer for downstream evaluation, and clinical usage scenarios.

2. Corpus Design

The GeMTeX corpus design workflow is displayed in Figure 1 and shows how raw clinical documents are taken from the local hospital information systems. Importantly, all contributing university hospitals are committed to processing only documents for which the patients actively signed the MII broad consent, i.e., they agreed to their data being used in research. Next, all documents undergo a manual

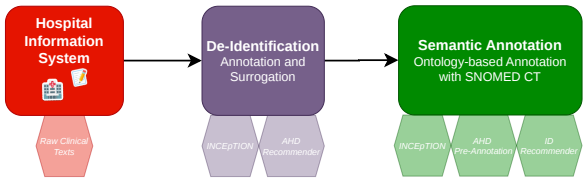


Figure 1: Schematic overview about the GeMTeX corpus design workflow. Clinical routine documents are taken from local hospital information systems, de-identified, and semantically annotated.

de-identification process, where relevant spans are annotated by two independent annotators using *INCEpTION* as annotation tool (Klie et al., 2018; Eckart De Castilho et al., 2024).[2] The annotation is supported by the industrial recommender system powered by *Averbis GmbH*[3]. A third person resolves disagreements before the annotated spans are replaced by pseudonyms[4] (Lohr et al., 2024, 2025).

Once de-identified, the documents undergo the final stage: The semantic annotation grounded in the widely shared and standardized ontology-based terminology Snomed CT in its April 2024 international version. To guide annotators through this complex task, Hofenbitzer et al. (2025) developed a comprehensive annotation guideline, which categorizes the 370k available Snomed CT concepts into three major groups and defined six annotation

[1]https://www.snomed.org/

[2]Access our de-identification guidelines and a gold standard dataset via https://doi.org/10.5281/zenodo.11502328.

[3]https://averbis.com/health-discovery/

[4]https://github.com/medizininformatik-initiative/GeMTeX/tree/main/surrogator

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 98–100. Palma, Mallorca. May 11, 2026.

maxims[5]

Each document annotation is performed by a single annotator with a medical background, supported by industrial pre-annotation from *Averbis GmbH* and a recommendation system powered by *ID Berlin*[6]. All annotators are employed as student assistants and receive fair compensation for their work. For quality assurance, 1% of the annotated documents are multiply annotated to compute site-level agreement and enable targeted adjudication. Periodic local and cross-site calibration sessions are instantiated to mitigate annotation drift.[7]

3. Current Status

The GeMTeX resource is in active implementation. As of February 2026, the project has delivered 15.4k de-identified documents (20.3M tokens, 682K de-identified spans), of which 382 documents (791.5K tokens) are semantically annotated, comprising 189.4K annotations. Besides the beforementioned annotation guidelines and gold standard examples, GeMTeX has designed a FHIR[8]-based interface for text material (Ammon et al., 2024).[9].

4. Accessibility

GeMTeX follows an access model under MII governance. Raw, de-identified, and semantically annotated texts remain at contributing sites and are not unconditionally redistributed. Access to the corpus or subsets is managed via the *German Portal for Medical Research Data* (FDPG)[10] under MII governance. Requests require a study protocol, ethics approval, and a data use application. The most recent information on accessibility options can be found at https://kiinformatik.mri.tum.de/en/gemtex, and interested researchers may reach out via gemtex.mi@mh.tum.de.

5. Outlook

The final project phase of GeMTeX focuses on completing annotations and consolidating guidelines, as well as releasing tooling. GeMTeX is intended to support benchmarking for named entity recognition, Snomed CT entity linking, as well as cross-site linguistic or medical analyses. By combining multi-site coverage, an ontology-grounded semantic layer, and controlled access, GeMTeX aims to provide a sustainable resource for clinical text-based research. In addition, structured clinical data, e.g., diagnoses, treatments, or laboratory values, are available for included patients and can be analyzed jointly with text data upon request.

6. Acknowledgments

We owe special thanks to all GeMTeX consortium members, associated partners, and contributing clinics at the university hospitals. This work is funded by the Federal German Ministry of Research, Technology, and Space under the grants 01ZZ2314A and 01ZZ2314B.

References

Danny Ammon, Maximilian Kurscheidt, Karoline Buckow, Toralf Kirsten, Matthias Löbe, Frank Meineke, Fabian Prasser, Julian Saß, Ulrich Sax, Sebastian Stäubert, Sylvia Thun, Reto Wettstein, Joshua P Wiedekopf, Judith A H Wodke, Martin Boeker, and Thomas Ganslandt. 2024. Arbeitsgruppe interoperabilität: Kerndatensatz und informationssysteme für integration und austausch von daten in der Medizininformatik-Initiative. 67(6):656–667.

Richard Eckart De Castilho, Jan-Christoph Klie, and Iryna Gurevych. 2024. Integrating INCEpTION into larger annotation processes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 110–121, Miami, Florida, USA. Association for Computational Linguistics.

Jakob Faller, Christina Lohr, Martin Boeker, and Frank Meineke. 2025. Building the Infrastructure for the German Medical Text Corpus Project (GeMTeX). *Studies in Health Technology and Informatics*, 327:894–895.

Udo Hahn. 2025. Clinical document corpora—real ones, translated and synthetic substitutes, and assorted domain proxies: a survey of diversity in corpus design, with focus on German text data. *JAMIA Open*, 8(3):ooaf024.

Justin Hofenbitzer, Stefan Schulz, Martin Boeker, Peter Klügl, Sarah Riepenhausen, Christina Lohr, Jacqueline Lammert, Andrea Riedel, and Luise Modersohn. 2025. Introducing Medical Semantic Annotation Guidelines for German Clinical Documentation with SNOMED CT.

[5] Access the semantic annotation guidelines via https://doi.org/10.5281/zenodo.15689930.

[6] https://www.id-berlin.de

[7] Access a single, synthetic semantic gold standard document via https://doi.org/10.5281/zenodo.18861607.

[8] https://www.hl7.org/fhir/

[9] https://www.medizininformatik-initiative.de/Kerndatensatz/KDS_Dokument/MIIIGModulDokument.html

[10] https://forschen-fuer-gesundheit.de/en/

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Christina Lohr, Jakob Faller, Andrea Riedel, Hung Manh Nguyen, Markus Wolfien, Justin Hofenbitzer, Luise Modersohn, Jutta Romberg, Fabian Prasser, Jazia Omeirat, Yutong Wen, Oksana Galusch, Udo Hahn, Marvin Seiferling, Christoph Dieterich, Peter Klügl, Franz Matthies, Janina Kind, Martin Boeker, Markus Löffler, and Frank Meineke. 2025. GeMTeX's De-Identification in Action: Lessons Learned & Devil's Details. In *German Medical Data Sciences 2025: GMDS Illuminates Health*, pages 274–282. IOS Press.

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*, volume 317, pages 171–179. IOS Press.

Frank Meineke, Luise Modersohn, Markus Loeffler, and Martin Boeker. 2023. Announcement of the German Medical Text Corpus Project (GeMTeX).

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana K. Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9:1–13, article no. 12.

Sebastian C Semler, Frank Wissing, and Ralf Heyder. 2018. German medical informatics initiative. *Methods of Information in Medicine*, 57(S 01):e50–e56.

From Corpus to Community: New NLP Tools for Welsh Language Research and Learning

Dawn Knight[1], Fernando Alva-Manchego[2]

[1]School of English, Communication and Philosophy; [2]School of Computer Science and Informatics, Cardiff University, UK

[1]KnightD5@cardiff.ac.uk, [2]AlvaManchegoF@cardiff.ac.uk

Abstract

Launched in 2020, CorCenCC (*Corpws Cenedlaethol Cymraeg Cyfoes* – National Corpus of Contemporary Welsh) is the first large-scale corpus of the Welsh language to integrate spoken, written, and electronically mediated data, offering a comprehensive snapshot of contemporary Welsh use. Including contributions from over 2,000 speakers, the 11.2-million-word corpus represents the diversity of Wales's linguistic landscape. As a national resource, CorCenCC enables users to explore real world Welsh. Several tools and resources were developed through the CorCenCC project, including the CyTag POS tagger and CySemTag (adapted from Lancaster University's USAS semantic system), to enable the grammatical and semantic categorisation of the dataset. The team also built the pedagogic toolkit *Y Tiwtiadur*, to allow learners and teachers to access corpus-based examples and tasks. Additionally, *Yr Amliadur* provides curated frequency-based wordlists across modes and parts of speech, supporting linguistic analysis and vocabulary development. Since completing the corpus, the team has focused on extending its impact and reach, to ensure that the resources are maintained and sustained for future use; a challenge often faced when large-scale projects end. This poster profiles the tools and resources created from and inspired by CorCenCC and its associated tools and resources, as a means of supporting the democratisation of linguistic resources for minoritised language contexts.

Keywords: CorCenCC, national corpus, NLP, Welsh, FreeTxt, *Proffiliadur, Y Tiwtiadur, Yr Amliadur*

1. Introduction

Launched in 2020, CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – National Corpus of Contemporary Welsh) is the first large-scale corpus of the Welsh language to integrate spoken, written, and electronically mediated data, offering a comprehensive snapshot of contemporary Welsh use. Including contributions from over 2,000 speakers, the 11.2-million-word corpus represents the diversity of Wales's linguistic landscape: regions, demographics, text types, modes, and genres. As a national resource, CorCenCC enables users to explore real-world Welsh, supporting research, teaching, lexicography, and translation (Knight et al, 2020a; Knight et al., 2020b). A screenshot of CorCenCC's query tools is provided in Figure 1.

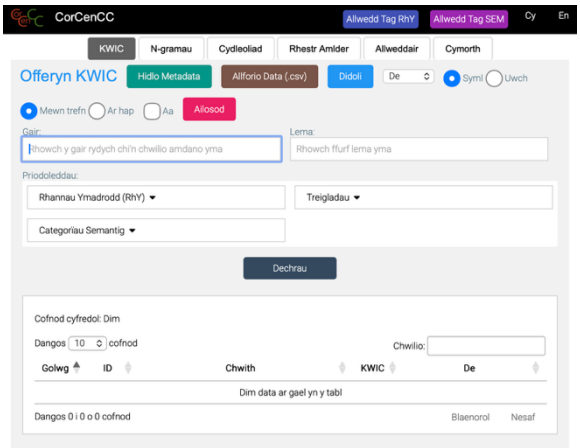


Figure 1: CorCenCC's query tools.

Several tools and resources were developed through the CorCenCC project, including the CyTag POS tagger (Neale et al., 2018) and CySemTag (adapted from Lancaster University's USAS semantic system – Piao et al., 2018), to enable the grammatical and semantic categorisation of the dataset. The team also built the pedagogic toolkit *Y Tiwtiadur*, to allow learners and teachers to access corpus-based examples and tasks. Figure 2 depicts the Word Identifier (*Abnadbod Geiriau*) functionality in the *Y Tiwtiadur* toolkit.

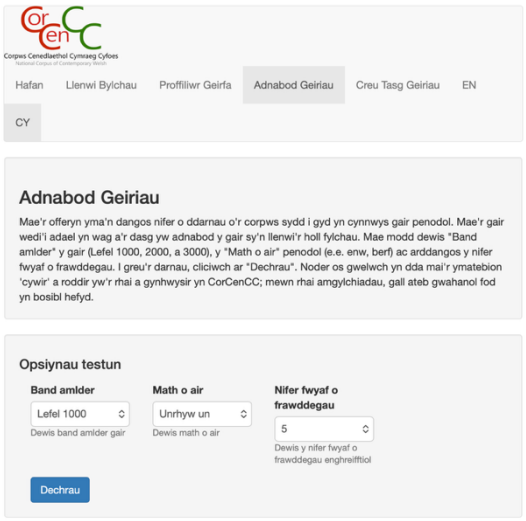


Figure 2: *Y Tiwtiadur*'s Word Identifier tool.

Additionally, *Yr Amliadur* provides curated frequency-based wordlists across modes and parts of speech, supporting linguistic analysis and vocabulary development (Knight et al. 2023).

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 101–103. Palma, Mallorca. May 11, 2026.

2. Extending CorCenCC's reach

Since completing the corpus, the team has focused on extending its impact and reach, to ensure that the resources are maintained and sustained for future use; a challenge often faced when large-scale projects end. This poster profiles the tools and resources created from and inspired by CorCenCC and its associated tools and resources, as a means of supporting the democratisation of linguistic resources for minoritised language contexts.

First, the poster profiles the development of *Geirfan*, a pedagogic wordlist created through a partnership with the National Centre for Learning Welsh, the Welsh Joint Education Committee, and language experts. Building on *Yr Amliadur*, the team developed frequency-driven vocabulary lists tailored to A1-level adult learners and created a prototype online dictionary. Since 2022, around 1,600 candidates have taken WJEC assessments based on *Geirfan* resources annually, marking the first time Welsh for Adults curricula have drawn directly on corpus-derived frequency data.

The poster also profiles the Welsh Government-funded ACC Welsh Automatic Text Summarisation tool (El-Haj et al., 2022; Ezeani et al., 2022), which allows users to generate concise summaries of long Welsh texts, supporting teaching and public access to information. Combining ACC, the CorCenCC dataset and its taggers, FreeTxt (Knight et al., 2024), is another resource developed by members of the team, as seen in Figure 3. FreeTxt enables Welsh and English qualitative data analysis using corpus-based NLP methods in an accessible interface co-designed with major Welsh cultural and educational organisations.

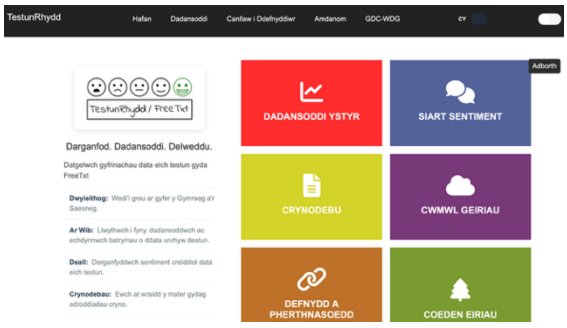


Figure 3: FreeTxt

Furthermore, *Proffiliadur,* our Python-based readability toolkit (Gutiérrez-Rolón et al., 2026), provides the first dedicated text-profiling resource for Welsh, offering reproducible, linguistically grounded measures of readability in a low-resource language.

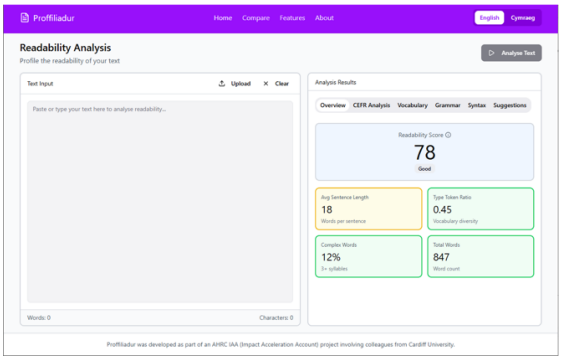


Figure 4: *Proffiliadur*

Proffiliadur computes 141 surface, lexical, morphological, and syntactic indices, designed to capture linguistic variation while incorporating Welsh-specific processing that enables accurate morphological analysis and handles phenomena such as initial consonant mutation. Proffiliadur enables assessment of text accessibility, supporting applications in education, healthcare, and public communication.

Finally, Gutierrez-Rolón and Alva-Manchego (2026) developed a mutation trigger identifier using, among other resources, the CyTag POS tagger and CorCenCC's query tool. The latter was instrumental in identifying examples of various types of Welsh mutations in real-world usage.

3. Sustainability and next steps

Through developing of these tools, the team has worked to empower end-users to direct and lead their own analyses of both small-scale and more extensive qualitative datasets to maximise the reach and potential impact. The approaches used to construct the resources serve as a template for those seeking to develop corpora and democratise language technology use in other minoritised and major language contexts around the world (e.g. Nguyen et al., 2026).

All resources are freely accessible via our Welsh Government funded DigiGrid platform (see Figure 4), which brings together a suite of tools to support Welsh language exploration, learning, and analysis.

Figure 5: The DigiGrid platform.

4. Acknowledgments

The work reported on in this poster is broadly based on the ESRC (Economic and Social Research Council) and AHRC funded Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction project (ES/M011348/1). The FreeTxt project was funded by AHRC (Arts and Humanities Research Council) follow-on funding for impact and engagement (AH/W004844/1). The development of the *Geirfan* wordlists and *Proffiliadur* toolkit were funded by Cardiff University's AHRC IAA account. Finally, the ACC Welsh Automatic Text Summarisation tool and DigiGrid platform were funded by Welsh Government's Welsh Language Technology grants.

5. Bibliographical References

El-Haj, M., Ezeani, I., Morris, J. and Knight, D. (2022). Creation of an evaluation corpus and baseline evaluation scores for Welsh text summarisation. *Proceedings of the Celtic Language Technology Workshop, LREC (Language Resources Evaluation) 2022 Conference*, June 2022, Marseille, France.

Ezeani, I., El-Haj, M., Morris, J., & Knight, D. (2022). Introducing the Welsh text summarisation dataset and baseline systems. *Proceedings of the LREC (Language Resources Evaluation) 2022 Conference*, June 2022, Marseille, France.

Gutiérrez-Rolón, N., Davies, J., Williams, T., Knight, D. & Alva-Manchego, F. (2026). Proffiliadur: Welsh Language Text Profiling Toolkit. *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC), May 2026, Palma de Mallorca, Spain.

Gutiérrez-Rolón, N. & Alva-Manchego, F. (2026). Unsupervised Labelling of Mutation Triggers in Welsh. *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC), May 2026, Palma de Mallorca, Spain.

Knight, D., Loizides, F., Neale, S. Anthony, L., & Spasić, I. (2020a). Developing computational infrastructure for the CorCenCC corpus – the National Corpus of Contemporary Welsh. *Language Resources and Evaluation, 55*(1)*,* 1-28.

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., Muralidaran, V., Tovey-Walsh, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M. and Scannell, K. (2020b). CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh. Cardiff University. http://doi.org/10.17035/d.2020.0119878310

Knight, D., Fitzpatrick, T., Morris, S., Tovey-Walsh, B., Prosser, H., & Davies, E. (2023). Corpus to curriculum: Developing word lists for adult learners of Welsh. *Applied Corpus Linguistics, 3*(2), article number: 100052.

Knight, D., Khallaf, N., El-Haj, M., Ezeani, I., & Morris, S. (2024). FreeTxt: a corpus-based bilingual free-text survey and questionnaire data analysis toolkit. *Applied Corpus Linguistics, 4*(3), article number: 100103.

Neale, S., Donnelly, K., Watkins, G., & Knight, D. (2018). Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh. *Proceedings of the LREC (Language Resources Evaluation) 2018 Conference*, May 2018, Miyazaki, Japan.

Nguyen, H. H., El-Haj, M., Rayson, P., & Knight, D. (2026). FreeTxt-Vi: A Benchmarked Vietnamese-English Toolkit for Segmentation, Sentiment, and Summarisation. Proceedings of *Learning Resources Evaluation Conference 2026* (LREC), May 2026, Palma de Mallorca, Spain.

Piao, S., Rayson, P., Knight, D., & Watkins, G. (2018). Towards a Welsh semantic annotation system. *Proceedings of the LREC (Language Resources Evaluation) 2018 Conference*, May 2018, Miyazaki, Japan.

Swiss-AL: Language Data Platform for Applied Sciences

Julia Krasselt, Philipp Dreesen, Dolores Lemmenmeier-Batinić, Sooyeon Geckeler, Klaus Rothenhäusler, Matthias Fluor

Zurich University of Applied Sciences, Institute of Language Competence
Theaterstrasse 17, 8400 Winterthur, Switzerland
{krss, dree, leme, chos, rotk, fluor}@zhaw.ch

Abstract

This paper introduces Swiss-AL, a language data platform designed for the multilingual, comparative analysis of public discourse in Switzerland. Swiss-AL is an open research data resource providing browser-based access to a variety of corpora in all four of Switzerland's official languages. Corpora contain journalistic, organisational, and parliamentary discourse. The platform supports research in applied linguistics as well as neighbouring disciplines (e.g., social sciences, communication and media studies).

Keywords: multilingual discourse corpus, public discourse, corpus analysis platform, applied sciences

1. Swiss-AL: Purpose and Composition of the resource

Swiss-AL is a language data platform for applied sciences, hosted and developed by the Digital Discourse Lab at ZHAW University of Applied Sciences (http://swiss-al.zhaw.ch). It is designed for the analysis and comparison of multilingual public discourses, with a focus on Switzerland. Swiss-AL is part of the Swiss linguistic research infrastructure landscape CLARIN-CH and a key component of the CLARIN knowledge centre for applied comparative discourse analysis (CLARIN-APPLIED, www.clarin-applied.zhaw.ch), hosted at ZHAW.

Swiss-AL contains a collection of linguistic corpora comprising publicly available documents from political, industrial, civil society, scientific and journalistic actors from all four language regions of Switzerland. Swiss-AL is not intended as a reference corpus, e.g. for Swiss High German, but rather as an empirical basis for analysing communicative practices in discursively constructed communication contexts. Swiss-AL is composed of three main corpus types:

- Journalistic corpora containing news articles published in Swiss daily and weekly newspapers and magazines by the country's leading publishing houses. The data covers a period from 2010 onwards and is available in all four Swiss national languages. It is provided by the Swiss Media Database via Swissdox@LiRi (provided by Zurich University).
- Organizational corpora containing press releases, news items, and blog posts from the official websites of over 360 actors in politics, administration, science, industry, and civil society. For example, these include the websites of all 26 Swiss cantons, the websites of parties represented in the Swiss National Council, and the websites of all Swiss universities. The data is available in three languages (German, French and Italian), covers the period from 2010 onwards, and is collected using a web-crawling and -scraping pipeline developed at ZHAW (Krasselt et al., 2023).
- Parliamentary corpora containing transcripts of speeches given by politicians in national parliamentary debates. The data is available in German, French and Italian, covers the period from 1999, and is provided via the parliamentary service's API.

In addition, Swiss-AL contains corpora compiled in the context of specific research projects conducted by the ZHAW Digital Discourse Lab (e.g, on covid-19 and vaccination discourses).

In the context of the Swiss Open Science Strategy, Swiss-AL has been developed into an Open Research Data resource since 2022 (Krasselt et al., 2023). This includes the implementation of FAIR-principles, the development and provision of a browser-based workbench for corpus analysis and the systematic consideration of legal issues concerning the publication of language data.

2. Access and Functionalities

A browser-based workbench is available for researchers to access and analyse the corpora. As the central interface for Swiss-AL, it provides access to corpus data in accordance with legal requirements, particularly with regard to copyright restrictions and data protection. This means that corpora cannot be downloaded and only short extracts of documents are displayed directly on the workbench. Where possible, a link to the original document on an external website is provided (e.g. Swissdox Essentials for all Journalistic corpus documents).

The workbench enables users to access a wide range of Swiss-AL corpora (see Table 1 for a selection) and create custom subcorpora based on criteria such as source, search terms and time spans. In the dedicated workspace, users can open multiple tabs and choose between the

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 104–105. Palma, Mallorca. May 11, 2026.

following data-driven and search-term-based analysis modes.

- In data-driven exploratory mode, users can analyse the corpus without entering a specific search term. This mode provides word embedding models and LDA-based topic models for all corpora (with predefined parameters such as number of topics), as well as keyword lists for all user-created subcorpora, currently based on Log Likelihood (Figure 1).

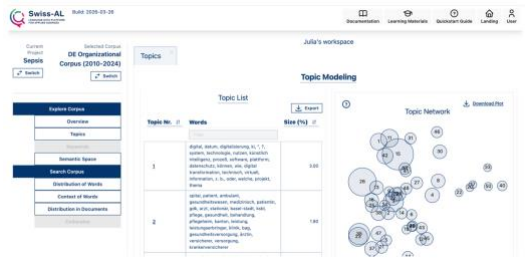


Figure 1: Screenshot of the topic modeling module on the Swiss-AL workbench

- In the keyword-based analysis mode, users can enter a specific search term and choose from three different search modes: simple, basic, or a CQP-based advanced search. Users can analyse the frequency and distributional patterns of a given search term, view classical and document-based concordances, calculate collocations (currently based on LogDice) and access the full text containing the search term on platform external websites (Figure 2).

	Corpus	Size (in token)
Journalistic Corpora	DE Journalistic Corpus (high reach media, 2010-2025)	1.2 billion
	IT Journalistic Corpus (high reach media, 2010-2025)	32 million
	FR Journalistic Corpus (high reach media, 2010-2024, 20%)	1.3 billion
	RM Journalistic Corpus (high reach+regional media, 2018-2025)	22 million
Organizational Corpora	DE Organizational Corpus (2010-2024)	66 million
	FR Organizational Corpus (2010-2024)	29 million
	IT Organizational Corpus (2010-2024)	18 million
Parliamentary Corpora	DE Swiss Federal Parliament Debates Corpus (1999-2024)	10 million
	FR Swiss Federal Parliament Debates Corpus (1999-2024)	4 million
	IT Swiss Federal Parliament Debates Corpus (1999-2024)	150,000

Table 1: Selection of corpora available on the Swiss-AL workbench (DE = German, FR = French, IT = Italian, RM = Romansh)

3. Bibliographical References

Krasselt, J., Dreesen, P., Fluor, M., & Rothenhäusler, K. (2023). Swiss-AL. Korpus und Workbench für mehrsprachige digitale Diskurse. In M. Kupietz & T. Schmidt (Eds.), *Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022*, pp. 127-142

Krasselt, J., Dreesen, P., Stücheli-Herlach, P., Lemmenmeier, D., Cho, S., Rothenhäusler, K., & Fluor, M. (2023). Swiss-AL: Platform for Language Data in Applied Sciences: On Challenges in the Field of Language Open Research Data. *Proceedings of the Conference on Research Data Infrastructure*, *1*. https://doi.org/10.52825/cordi.v1i.249

EuReCo, KorAP and DeReKo: Updates on Ingestion and Annotation Pipelines, Backend, Interfaces, Operation, and Corpora

Marc Kupietz, Harald Lüngen, Nils Diewald, Helge Stallkamp, Uyen-Nhu Tran, Rameela Yaddehige
Leibniz Institute for the German Language (IDS)
Mannheim, Germany
{kupietz, luengen, diewald, stallkamp, tran, yaddehige}@ids-mannheim.de

Abstract

This paper reports on recent technical developments in the European Reference Corpus EuReCo and its current technical implementation based on the corpus search and analysis platform KorAP. We describe updates to the ingestion pipeline, including extensions to the TEI-to-KorAP-XML converter `tei2korapxml` and the KorAP tokenizer, as well as the newly introduced `korapxmltool` for annotation and index conversion. We further present *KoralMapper*, a service that enables cross-schema comparability of annotations and metadata at query time, and report on developments in the backend access control system Kustvakt, the web user interface Kalamar, API client libraries for R and Python that promote reproducibility and methodologically sound AI-assisted analysis, and containerized deployment. The corpora and languages currently represented in EuReCo are outlined, and the role of the German Reference Corpus DeReKo, including its metadata-driven virtual corpus design, predefined useful subcorpora, and I5/TEI encoding, is discussed in detail.

Keywords: Comparable Corpora, Reference Corpora, Tokenization, Annotation, Interoperability, Containerization

1. Introduction

The European Reference Corpus EuReCo (Kupietz et al., 2017, 2024) is an open long-term initiative aimed at providing and using virtual, dynamically definable comparable corpora based on existing national reference corpora. Unlike parallel corpora, which are susceptible to shining-through effects, or web-based comparable corpora, which are limited to web material and typically lack rich metadata, EuReCo draws on large, carefully curated national corpora and thus offers a complementary approach to cross-linguistic research. This paper reports on recent, mostly technical developments underpinning EuReCo's implementation: the corpus search and analysis platform KorAP serves as EuReCo's current technical basis and is the focus of Sections 2, covering the ingestion pipeline, annotation, cross-schema comparability, backend, and deployment. Section 3.1 describes the role of DeReKo, the German Reference Corpus, as a major contributing corpus within EuReCo.

2. KorAP

KorAP (Bański et al., 2012; Diewald et al., 2016) is the corpus search and analysis platform that currently serves as the technical basis of EuReCo. Originally developed at IDS Mannheim primarily as an access point to DeReKo, KorAP takes an agnostic approach to data and research questions, rendering it applicable to corpora with arbitrary metadata and annotation schemes[1]. It supports an extensible set of query languages, localization, and a plugin architecture, and is openly developed under a BSD-2-clause license.[2] The following subsections report on recent developments in KorAP, covering the ingestion pipeline, annotation and metadata comparability, backend infrastructure, user interface, and operation.

2.1. Ingestion Pipeline

2.1.1. TEI Import: tei2korapxml

The fundamental layer where the ingestion pipeline is initiated is the import of TEI XML files. The TEI format is a widely used standard for encoding texts, and it allows for rich metadata and structural information to be included. The import process involves parsing the TEI XML files, extracting relevant information such as the text content (which is also tokenized), metadata (e.g., author, title, publication date), and structural elements (e.g., chapters, paragraphs). The result of the conversion is in KorAP XML format (Bański and Diewald, 2025), a radical standoff format, specifically designed for use with the KorAP corpus analysis platform. This format allows for efficient storage and retrieval of the text and its associated metadata, as well as for the parallel addition of various stand-off annotations in subsequent steps of the pipeline.

The tei2korapxml tool was originally designed to specifically handle the I5 customization (Lüngen

[1] Restrictions may concern, e.g., word segmentations.
[2] https://github.com/KorAP/

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 106–112. Palma, Mallorca. May 11, 2026.

and Sperberg-McQueen, 2012) of TEI P5. Since version 2.4 (as of February 2023), however, it has been extended to support general TEI P5 documents.

2.1.2. Tokenization

A very important part of the TEI import is the tokenization of the text content. The tei2korapxml tool can use any tokenizer, but recommends the use of the integrated KorAP tokenizer and sentence splitter (Diewald et al., 2022)[3], which has a highly efficient DFA, with a throughput of 10MB/s, as its core (based on JFlex), and has a comprehensive abbreviation list for German and abbreviation lists for other selectable languages.

Recent updates to the KorAP tokenizer include several bug fixes (e.g., soft hyphens are no longer handled as token boundaries), an update to support Unicode 15.0 (including emojis with zero-width joiners and skin-tone modifiers), extensions concerning frequent German abbreviations, and, since version 2.4, support for German gender-sensitive spelling. The latter now also includes the handling of colons, slashes, and brackets as separators for gender endings, in addition to the previously supported asterisk and underscore. Thus, nouns like *Lehrer:in*, adjectives like *schön:es*, and pronouns like *diese(r)* are now correctly tokenized as single tokens. To allow for reproducibility and the tokenization of older texts, the support for gender-sensitive spellings can be turned off (by selecting *de-old* as the language option).

2.1.3. Annotations

A key feature of the KorAP XML format is that it allows for the addition of arbitrary stand-off annotations, which go to separate KorAP XML ZIP files in subsequent steps of the pipeline. This means that after an initial conversion of the base TEI, possibly already including linguistic annotations, various, possibly multiple layers of annotations can be added to the text without modifying the original text content.

To add annotations, the new korapxmltool[4] can be used, which has already integrated support for Java based tools like OpenNLP[5], CoreNLP (Manning et al., 2014), MarMot (Mueller et al., 2013), and Malt-Parser (Nivre et al., 2006). In addition, via corresponding Docker containers provided by the KorAP team[6], support for TreeTagger (Schmid, 1994) and spaCy (Honnibal et al., 2020) is also integrated. Furthermore, any annotation tool that reads and writes the CoNLL-U format (Nivre et al., 2016) can be integrated.

2.1.4. KorAP XML to Krill Conversion

A central step in the ingestion pipeline is the merging of the various KorAP XML files — covering primary text, structure, and annotations — into a single Krill-JSON file. These files are encoded in JSON and serve as the basis for building the KorAP search index. In the original Perl-based prototype, this merging step constituted a significant bottleneck: it required unpacking often millions of individual XML files to the file system.

The rewritten Kotlin/Java implementation, now integrated into `korapxmltool`, eliminates this bottleneck through two key improvements: files are processed directly from their compressed form without intermediate unpacking, and concurrent shared-memory hashes are used to collect and merge information across files in parallel. The result is a speed-up of one to three orders of magnitude depending on text size, available CPU cores, and memory. As a concrete illustration, converting all German Wikipedia articles now takes approximately three hours rather than three weeks, a reduction that makes previously impractical large-scale re-indexing workflows feasible.

2.2. Comparability

While virtual corpora already enable comparability at the corpus level and various query languages, client software etc. support comparability at the analysis level in KorAP, there has been no mechanism for comparability at the annotation and metadata level, which is important for EuReCo research. *Koral-Mapper*[7] closes this gap; it is a service for the KorAP search platform that translates annotations and metadata between different schemas by rewriting the JSON-based intermediate representation KoralQuery (Bingel and Diewald, 2015) based on predefined transformation rules – both for requests and responses. This allows users to search, for example, for linguistic structures using Universal Dependency POS annotations even though only STTS tags are available in the index. The reverse transformation enriches the results accordingly. Since annotation and metadata schemas rarely have 1:1 relationships, the rules also support Boolean logic. The service uses a rewrite mechanism and can be integrated directly into the KorAP frontend, as well as via the API and client libraries.

[3] https://github.com/KorAP/KorAP-Tokenizer

[4] https://github.com/KorAP/korapxmltool

[5] https://opennlp.apache.org/

[6] https://hub.docker.com/repositories/korap?search=conllu

[7] https://github.com/KorAP/Koral-Mapper

2.3. The Web User Interface

KorAP is based on a modular client–server architecture. Its web-based user interface, Kalamar (Diewald et al., 2019), communicates with KorAP via the backend's publicly accessible web services (Kupietz et al., 2022). KorAP's principle of designing small, independent components (Diewald et al., 2016) is also reflected in its plugin mechanism. Plugins can be developed independently of the Kalamar interface and integrated into the Kalamar web application. They retrieve data via KorAP's web service API. This approach allows the Koral-Mapper service to be integrated into the Kalamar frontend as a plugin.

Beyond its modular architecture, Kalamar has been revised to improve usability and visual clarity (see Diewald et al., 2025). In version 0.60.0 the navigation structure has been redesigned to create a more modern and intuitive user experience, with an improved visual appearance of the navigation components and a more logical grouping of related functionalities. Kalamar follows a design approach based on "progressive disclosure", which ensures that basic functionalities are easily accessible while advanced options are only available on demand (Tidwell, 2006). The revised user interface improves clarity and reduces cognitive load by structuring visual elements according to the principles of proximity and consistency (Lidwell et al., 2010). A new top navigation bar now groups core functionalities such as login/logout, documentation, and other related items. In addition, the responsive layout has been revised to adapt the new navigation and improve accessibility across devices and user groups.

2.4. Backend

Kustvakt[8] operates the backend of KorAP, connects other components and manages their tasks. Importantly, it is responsible for user rights and access control management that aims at maximizing user access to corpus data while protecting rights holders' legitimate interests (Margaretha Illig et al., 2025). Kustvakt provides web service APIs to access virtual corpora and its various annotations, that mostly involve complex licenses and diverse restrictions depending on access methods and purposes.

KorAP's access policies are defined to model license forms, namely which users have which access rights to which data. Kustvakt enforces the access policies through query rewriting techniques (Bański et al., 2014). For instance, unauthorized requests are permitted to search only on free resources or the metadata of all resources, including protected ones. Supporting OAuth2 (Hardt, 2012), Kustvakt improves access capabilities and enables authorized access through third party applications. Kustvakt is open-source, extensible, and generally applicable for access control in corpus analysis tools.

2.5. API Client Libraries

Client libraries for R (Kupietz et al., 2020) and Python (Kupietz et al., 2022) facilitate access to KorAP's REST API, notably promoting reproducibility and replicability in research workflows. Since July 2025, their CI pipeline has included tests to ensure that LLMs can solve basic analysis tasks when prompted with the documentation. This doc-prompting approach aims to support "vibe-coding" analysis scripts in a way that is as methodologically sound and sustainable as possible, by ensuring the documentation remains a reliable basis for AI-assisted generation (Trawiński et al., 2025).

The client libraries can be used for all corpora accessible via KorAP, including those within EuReCo. Recent updates added support for comparing collocation analyses across multiple virtual corpora, an important feature for defining comparable corpora, to be used, e.g., for analyzing light verb constructions across languages and other distributional phenomena that are sensitive to topic domain and genre composition (Kupietz and Trawiński, 2022).

2.6. Operation and Containerization

KorAP is based on different independent components that can be installed and run via a single Docker Compose command.[9] An instance is based on a specific index and around 20 instances have been deployed at the IDS and more outside. Monitoring and managing multiple instances presented challenges, leading us to utilize Portainer[10] for central management of Docker containers.

3. Corpora and Languages represented in EuReCo

EuReCo is designed to be a long-term initiative that can be extended with new corpora and languages. Currently, EuReCo provides access, to varying degrees, to the following national and reference corpora: the German Reference Corpus DeReKo (Kupietz et al., 2010), the Contemporary Corpus of the Romanian Language CoRoLa (Barbu Mititelu et al., 2018, Romanian Academy, 2017), the Hungarian National Corpus HNC (Váradi, 2002; Oravecz et al., 2014, Hungarian Academy of Sciences, 2018), the Bulgarian National Reference

[8] https://github.com/KorAP/Kustvakt

[9] https://github.com/KorAP/KorAP-Docker

[10] https://www.portainer.io/

Corpus BNRC (Simov et al., 2004), the Polish Reference Corpus NKJP (Przepiórkowski et al., 2004).

3.1. DeReKo

The German Reference Corpus DeReKo is the largest text archive for linguistic research of contemporary German. DeReKo is constantly being extended; currently (as of DeReKo-2026-I), it contains 63.8 billion tokens (Leibniz-Institut für Deutsche Sprache, 2026). The composition of DeReKo, as illustrated in Table 1 with examples of widely used virtual subcorpora, covers a wide range of genres. The bulk of DeReKo has always consisted of press corpora, but it also contains fiction, specialized texts, Computer-mediated communication (CMC), and many other genres. Recent additions include paraliterature, a corpus of YouTube comments (Cotgrove, 2023; Kupietz et al., 2023), and journals specialized in engineering and technology (Lüngen et al., 2025).

Category	Tokens
Press	19,431,351,555
CMC (Wikipedia Talk, Usenet)	786,760,495
Plenary protocols	379,344,831
Fiction (Literature, Novels)	89,912,729
General-interest magazines	94,000,352
Specialized (Science, IT, etc.)	55,542,376

Table 1: Composition of DeReKo-KorAP-2026-I (the virtual subcorpus of DeReKo-2026-I that is available via KorAP) by selected predefined useful subcorpora.

DeReKo is a general-purpose corpus and adheres to a primordial sample design, which means that users define virtual subcorpora that are tailored to (e.g. balanced w.r.t.) a specific research question using DeReKo's metadata. Recently, we have published a list of "useful virtual corpora" (VCs) and their definitions via regular expressions over KorAP metadata fields. Each definition is linked to a webpage with a KorAP interface where the respective definition is specified in the corpus assistant, i.e. the user can immediately start with queries to the VC. Examples of such useful virtual subcorpora are *newspaper commentaries*, *novels*, or *IT magazines*.[11]

DeReKo is a major part of EuReCo and has been used in various pilot studies using comparable corpora, e.g. (Bański et al., 2023). (Pairs of) comparable corpora are designed by defining a mapping between relevant metadata of DeReKo and another corpus of a different language within EuReCo. DeReKo's rich metadata such as text type, topic domain, article type or newspaper column facilitate its use in EuReCo.

The basic text structure and text and corpus metadata are encoded in I5, a TEI customization developed for DeReKo (Lüngen and Sperberg-McQueen, 2012; Lüngen and Pisetta, 2025)

4. Summary and Conclusions

We have presented recent developments in the EuReCo initiative and its technical infrastructure. On the ingestion side, the `tei2korapxml` converter now supports general TEI P5 in addition to the I5 customization, and the KorAP tokenizer has been extended with full Unicode 15.0 support and handling of gender-sensitive spellings. The newly introduced `korapxmltool` consolidates annotation integration and KorAP XML-to-Krill conversion, achieving speed-ups of one to three orders of magnitude for the latter. With *Koral-Mapper*, cross-schema comparability of annotations and metadata is now possible at query time, a key prerequisite for EuReCo's comparative research agenda. On the backend, Kustvakt ensures fine-grained access control, while the revised Kalamar interface improves usability through progressive disclosure and a modernized navigation. Client libraries for R and Python promote reproducibility and support methodologically sound AI-assisted analysis through CI-tested documentation. Containerized deployment via Docker and centralized management with Portainer simplify the operation of multiple KorAP instances.

On the corpus side, DeReKo continues to grow and now comprises 63.8 billion tokens across a wide range of genres. The introduction of predefined useful virtual subcorpora lowers the barrier for researchers to work with well-defined subsets of the data.

Future work will focus on extending EuReCo's language and corpus coverage, further developing the Koral-Mapper rule sets for additional annotation schemas, and continuing to improve the sustainability of AI-assisted research workflows through the client libraries.

5. Bibliographical References

Verginica Barbu Mititelu, Dan Tufiş, and Elena Irimia. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[11] https://korap.ids-mannheim.de/doc/corpus/useful_subcorpora

Piotr Bański and Nils Diewald. 2025. Dealing with multiple annotations. In Piotr Bański, Ulrich Heid, and Laura Herzberg, editors, *Harmonizing language data. Standards for linguistic resources*, volume 4 of *Digital Linguistics*, pages 169–200. De Gruyter.

Piotr Bański, Nils Diewald, Michael Hanl, Marc Kupietz, and Andreas Witt. 2014. Access Control by Query Rewriting: the Case of KorAP. In *Proceedings of the 9th conference on the Language Resources and Evaluation Conference (LREC '14)*, pages 3817–3822, Reykjavic, Iceland.

Piotr Bański, Nils Diewald, Marc Kupietz, and Beata Trawiński. 2023. Applying the newly extended European reference corpus EuReCo. Pilot studies of light-verb constructions in German, Romanian, Hungarian and Polish. In *Book of Abstracts of the 10th International Contrastive Linguistics Conference (ICLC-10), 18-21 July, 2023, Mannheim, Germany*, pages 274–276, Mannheim. IDS-Verlag.

Piotr Bański, Peter M. Fischer, Elena Frick, Erik Ketzan, Marc Kupietz, Carsten Schnober, Oliver Schonefeld, and Andreas Witt. 2012. The New IDS Corpus Analysis Platform: Challenges and Prospects. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2905–2911, Istanbul, Turkey. European Language Resources Association (ELRA).

Joachim Bingel and Nils Diewald. 2015. KoralQuery – a General Corpus Query Protocol. In *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*, pages 1–5, Vilnius, Lithuania.

Louis Cotgrove. 2023. THE NOTTDEUYTSCH CORPUS: A corpus of German-language YouTube comments. *Korpora Deutsch als Fremdsprache*, 3(2). Number: 2.

Nils Diewald, Verginica Barbu Mititelu, and Marc Kupietz. 2019. The KorAP user interface. Accessing CoRoLa via KorAP. *On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*, 64(3). Place: Bucharest, Romania.

Nils Diewald, Franck Bodmer, Peter M. Fischer, Elena Frick, Marc Kupietz, Mark-Christoph Müller, Helge Stallkamp, and Uyen-Nhu Tran. 2025. Linguistic corpus research software at the Leibniz-Institute for the German Language (IDS). In *Postproceedings of the deRSE 2025*, number 85 in Electronic Communications of the European Association for Software Science and Technology, Berlin. Berlin Universities Publishing / deRSE. Status: toBePublished.

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP Architecture Diving in the Deep Sea of Corpus Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3586–3591, Portorož, Slovenia. European Language Resources Association (ELRA).

Nils Diewald, Marc Kupietz, and Harald Lüngen. 2022. Tokenizing on scale. Preprocessing large text corpora on the lexical and sentence level. Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany, pages 208 – 221. IDS-Verlag, Mannheim.

Dick Hardt. 2012. The OAuth 2.0 Authorization Framework. Request for Comments RFC 6749, Internet Engineering Task Force.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Marc Kupietz, Piotr Banski, Nils Diewald, Beata Trawinski, and Andreas Witt. 2024. EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 94–103, Torino, Italia. ELRA and ICCL.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German Reference Corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta, Malta. European Language Resources Association (ELRA).

Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2020. RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo via KorAP. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, pages 7015–7021, Marseille, France. European Language Resources Association.

Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2022. Building paths to corpus data: A multi-level least effort and maximum return approach. In Darja Fišer and Andreas Witt, editors, *CLARIN. The Infrastructure for Language Resources.*,

pages 163–189. deGruyter, Berlin. Section: number x.

Marc Kupietz, Harald Lüngen, and Nils Diewald. 2023. Das Gesamtkonzept des Deutschen Referenzkorpus DeReKo: Vom Design bis zur Verwendung und darüber hinaus. In Arnulf Deppermann, Christian Fandrych, Marc Kupietz, and Thomas Schmidt, editors, *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial*, pages 1–28. De Gruyter.

Marc Kupietz and Beata Trawiński. 2022. Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo. In Laura Auteri, Natascia Barrale, Arianna Di Bella, and Sabine Hoffmann, editors, *Wege der Germanistik in transkultureller Perspektive. Akten des XIV. Kongresses der Internationalen Vereinigung für Germanistik (IVG) (Bd. 6)*, Jahrbuch für Internationale Germanistik - Beihefte - 6, pages 417–439. Peter Lang, Bern.

Marc Kupietz, Andreas Witt, Piotr Bański, Dan Tufiş, Dan Cristea, and Tamás Váradi. 2017. EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017*, pages 15–19, Mannheim. Institut für Deutsche Sprache.

William Lidwell, Kritina Holden, and Jill Butler. 2010. *Universal Principles of Design*. Rockport Publishers, Beverly, Massachusetts.

Harald Lüngen, Marc Kupietz, Nils Diewald, and Helge Stallkamp. 2025. Potenziale der Gingko-Integration in DeReKo: Analyse mit KorAP, nachhaltige Verfügbarkeit und mehr. In Christian Fandrych, Annette Portmann, Lars Schirrmeister, and Franziska Wallner, editors, *„Weichgeglüht und luftvergütet". Potenziale eines ingenieurwissenschaftlichen Korpus für Forschung und Vermittlung*, volume 20 of *Deutsch als Fremd- und Zweitsprache. Schriften des Herder-Instituts (SHI)*, pages 89–112. Stauffenburg, Tübingen.

Harald Lüngen and Ines Pisetta. 2025. Conversion into the archival format I5. In Piotr Bański, Ulrich Heid, and Laura Herzberg, editors, *Harmonizing language data. Standards for linguistic resources*, volume 4 of *Digital Linguistics*, pages 229–250. De Gruyter.

Harald Lüngen and C. Michael Sperberg-McQueen. 2012. A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, 3.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McCloskey. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Eliza Margaretha Illig, Nils Diewald, Paweł Kamocki, and Marc Kupietz. 2025. Managing Access to Language Resources in a Corpus Analysis Platform. In *Proceedings of: Selected papers from the CLARIN Annual Conference 2024. Barcelona, Spain, 15–17 October 2024 (= Linköping Electronic Conference Proceedings 216).*, pages 101–112. Linköping University Electronic Press.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1667, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, pages 1719–1723, Reykjavik, Iceland. European Language Resources Association (ELRA).

Adam Przepiórkowski, Zygmunt Krynicki, ukasz Dębowski, Marcin Woliński, Daniel Janus, and Piotr Bański. 2004. A Search Tool for Corpora with Positional Tagsets and Ambiguities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*,

pages 1235–1238, Lisbon, Portugal. European Language Resources Association (ELRA).

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, and Dimitar Doikoff. 2004. A Language Resources Infrastructure for Bulgarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Jenifer Tidwell. 2006. *Designing Interfaces: Patterns for Interaction Design*. O'Reilly.

Beata Trawiński, Marc Kupietz, and Nils Diewald. 2025. News from EuReCo: Annotations, Applications, and LLM Assistance. page 3, Karlova. Filozofická Fakulta Univerzita Karlova.

Tamás Váradi. 2002. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 385–389, Las Palmas, Spain. European Language Resources Association (ELRA).

6. Language Resource References

Hungarian Academy of Sciences. 2018. *Hungarian National Corpus*.

Leibniz-Institut für Deutsche Sprache. 2026. *DeReKo-2026-I*. Leibniz-Institut für Deutsche Sprache, German Reference Corpus DeReKo, DeReKo-2026-I.

Romanian Academy. 2017. *Reference Corpus of Contemporary Romanian Language*. Romanian Academy, CoRoLa.

Author Index