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. *©ELRA Language Resources Association, 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.