Third Austrian Meeting on Digital Linguistics /

Drittes Österreichisches Treffen zur Digitalen Linguistik

BOOK OF ABSTRACTS

49. Österreichische Linguistiktagung 2025 (ÖLT 2025)

December 7, 2025

Third Austrian Meeting on Digital Linguistics @ ÖLT 2025: Introduction

Digital linguistics is a growing interdisciplinary field at the intersection of linguistics, information technology, and the social sciences. This is reflected by a growing number of new projects, publication series, and university courses. A central focus of digital linguistics is language data, i.e., digital artifacts that use human language as a form of expression. The range of these language data includes social media content, parliamentary transcripts, newspapers, and medieval manuscripts, among others. Such data are processed, annotated, analyzed, curated, shared, archived, and reused, among other activities. Also new technologies such as large language models (LLMs) and generative AI play a growing role in digital linguistics. Therefore, the topics covered in this workshop span from the creation of digital language resources (corpora, dictionaries, etc.), new methods (application of LLMs and generative AI), analysis of language data (e.g., semantic change detection, emotion and sentiment analysis), to the use of standards and research infrastructures, as well as methods for long-term archiving or reuse of language data.

The variety of research in this field in Austria was shown during the first Austrian Meeting on Digital Linguistics and the second Austrian meeting on Digital Linguistics, as well as within the context of the previous Austrian Meeting on Sentiment Inference (ÖTSI 2021, 2023), where 37 researchers from different Austrian and international research institutions presented their projects.

This year's workshop "Third Austrian Meeting on Digital Linguistics" is a continuation of this workshop series, organized in the framework of CLARIAH-AT. Again, the aim of the workshop is to highlight recent developments in the Austrian research landscape and to connect different projects working with or on methods in digital linguistics, as well as the researchers involved. The workshop aims to facilitate the exchange of methodological insights and the creation of synergies through the mutual sharing of digital language resources, also within the framework of the research infrastructure CLARIAH-AT. Furthermore, the workshop also addresses international researchers, who are working in the field of digital linguistics and who want to present their research and exchange and connect with the Austrian research community.

The workshop programme is composed of 14 presentations, which were peer-reviewed.

The workshop is supported by CLARIAH-AT.

Workshop organizers: Tanja Wissik, Andreas Baumann, Julia Neidhardt, Claudia Posch, Gerhard Rampl

Programme

Third Austrian Meeting on Digital Linguistics (Sonntag, 7. Dezember 2025)	
Uhrzeit	
Raumnummer ÖLT	9
Raumnummer (Uni)	N.0.27 (30)
9:00-9:05	Tanja Wissik, Julia Neidhardt, Andreas Baumann "Introduction"
9:05-9:30	Tanja Wissik, Maciej Ogrodniczuk, Petya Osenova, "Interoperable Corpora of Historical Newspapers: the PressMint Project
9:30-10:00	Klara Venglerova "Processing Digitized Text on an Example of Job Advertisements from Austrian Periodicals from 1850-1950"
10:00-10:30	Jona Marie Hassenbach, Magdalena Miteva "Potential of Generative AI for Text Transcription"
10:30-11:00	
11:00-11:30	Teodor Petrič, "Large language models as synthetic participants in psycholinguistic experiments: the case of German noun plural formation.
11:30-12:00	Varvara Arzt, Allan Hanbury, Terra Blevins, "Analysis of Word Order Biases in Language Models: A Controlled Investigation Using Artificial and Natural Languages"
12:00-12:15	Natalia Borza, "The reliability of detecting and exploring basic emotions in short social media texts using the BEMDI-metre" (short paper)
12:15-12:30	Edlira Gugu, "Phylogenetic analysis of folktale evolution: Reconstructing cultural transmission through computational Methods" (short paper)
12:30-14:30	
14:30-15:00	Michelle van de Bilt, Florian Jung, Matthis Hupertz, Irene Böhm, Nikolaus Ritt, "The rise of present participial -ing in Middle English: Reducing ambiguity in word structure Signals"
15:00-15:30	Cordula Meißner, Janina Deilke, Anna-Lena Randermann, "Kommunikationsverben als Indikatoren für situativen Kontext – Ein Forschungsprojekt zu Indexikalität in Korpusanalyse und Sprachwissen"
15:30-16:00	Rashid Mustafin, "Philosophical considerations of digital text analysis: Core assumptions and methodological challenges"
16:00-16:30	
16:30-16:45	Katharina Horn, Fabian Navarro, Jasmin Bettstein, "Vergleichende Analyse von nominierten Texten des Ingeborg-Bachmann-Preises" (short paper)
16:45-17:00	Ilia Afanasev "Heatmap-based visualisation of the linguistic variation (on the material of East Slavic small territorial lects)" (short paper)
17:00-17:30	Nataliia Cheilytko, "Regional Semantic Change and Variation in Ukrainian with LLMs"
17:30-18:00	Juliane Benson, Julia Neidhardt, Katharina Zeh, Andreas Baumann, Hannes Essfores, Hannes Fellner, "Linguistic diversity and digitalization: a progress report

Heatmap-based visualisation of the linguistic variation (on the material of East Slavic small territorial lects)

Ilia Afanasev

In modern linguistics, the visualisation techniques are often used only as a way to represent research results, lacking any kind theoretical depth (cf. the critique of tree-based visualisation in phylogenetic linguistics by List et al. (2014: 203-204)). This study instead advocates the use of visualization techniques (Unwin, 2024) as an exploratory tool for uncovering patterns identified through manual tagging. The utilised approach applies a heatmapbased representation of the various (phonological, morphological, lexical and syntactical) levels of language variation in small territorial lect (dialect) texts to show the differences between their distributions (or lack thereof). The main research dataset comprises material from various historically attested East Slavic dialects, predominantly Transcarpathian Slavic (Bojkian, Lemkian, and Huzulian) (Nakonečna and Rudnyc'kyj, 1940). The collected small corpus (approximately 10,000 tokens) is subjected to a preliminary manual analysis informed by existing research on the historical development of the Slavic clade (among others, Hancov (1974); Zhovtobriukh et al. (1979); Holzer (1995)). For comparative purposes, the study draws on data from modern East Slavic dialects, primarily the Saratov dialectological corpus (Kryuchkova and Goldin, 2011) and the Khislavichi corpus (Ryko and Spiricheva, 2020). The study provides additional statistical and qualitative analysis to compare the distributions within the data from a linguistic perspective. The study offers a new perspective on onomasiological reconstruction and corpus material dynamics in historical comparative linguistics, while showing the benefits of interaction between software engineering and computational linguistics, as well as the advantages of enhancing quantitative methods with qualitative analysis. The software, developed during the course of the study, is made open-access to ensure the reproducibility of the research in accordance with current practices.

References

Hancov, V. (1974), Dijalektolohična klasyfikacija ukrajinśkych hovoriv, [nachdr. der ausg.] kyïv, 1923 edn, Böhlau, Köln.

Holzer, G. (1995), 'Die einheitlichkeit des slavischen um 600 n. chr. und ihr zerfall', Wiener Slavistisches Jahrbuch 41, 55–89. URL: http://www.jstor.org/stable/24748655

Kryuchkova, O. Y. and Goldin, V. E. (2011), Korpus russkoj dialektnoj rechi: koncepcija i parametry ocenki [a corpus of russian dialectal speech: the concept and parameters of evaluation], in 'Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2011)', Vol. 10, Russian State University for the

Humanities, pp. 359–367.

List, J.-M., Shijulal, N.-S., Martin, W. and Geisler, H. (2014), 'Using phylogenetic networks to model chinese dialect history', Language dynamics and change 4(2), 222–252.

Nakonečna, H. and Rudnyc'kyj, J. B. (1940), Ukrainische Mundarten: Südkarpatoukrainisch ; (Lemkisch, Bojkisch und Huzulisch), Arbeiten aus dem Institut für Lautforschung an der Universität Berlin ; 9, Otto Harrassowitz, Berlin.

Ryko, A. I. and Spiricheva, M. V. (2020), ‘Corpus of the russian dialect spoken in khislavichi district’. URL: https://lingconlab.ru/khislavichi/

Unwin, A. (2024), Getting (More out of) Graphics: Practice and Principles of Data Visualisation, Chapman and Hall/CRC. URL: https://www.routledge.com/Getting-more-out-of-Graphics-Practice-and-Principles-of-Data-Visualisation/Unwin/p/book/9780367673994

Zhovtobriukh, M. A., Rusanivs’kyˇı, V. M. and Skliarenko, V. H. (1979), Istorija ukraïns’koï movy. Fonetyka: [monohrafija], Naukova dumka, Kyïv.

Analysis of Word Order Biases in Language Models: A Controlled Investigation Using Artificial and Natural Languages

Varvara Arzt, Allan Hanbury, Terra Blevins

This study investigates word order biases in transformer-based language models (Vaswani et al. 2017), motivated by concerns that such biases may impact linguistic diversity in the future (Anderson et al. 2024; Guo, Conia, et al. 2025; Guo, Shang, et al. 2024). We examine how auto-regressive transformer-based models develop certain word order preferences through both inductive and distribution biases, addressing whether these preferences reflect linguistic universals (Greenberg 1963) or model-specific artefacts. Our methodology employs controlled parallel experiments with artificial and natural languages using identical model architectures. Artificial languages, generated via probabilistic context-free grammars (Booth & Thompson 1973; Chomsky 1981), isolate word order variables by systematically varying constituent arrangements whilst controlling for other linguistic factors (White & Cotterell 2021). To test robustness of our findings, we perform identical experiments with typologically similar natural languages. Beyond that, experiments with natural languages employ both monolingual and multilingual models to assess whether multilingual models exhibit systematic biases towards dominant training languages (Guo, Conia, et al. 2025). We address three core questions: first, whether transformers demonstrate systematic preferences for specific word orders; second, how acquisition efficiency varies across languages with varying word order; and third, whether model preferences align with established typological correlations, such as the relationship between basic word order and adposition placement documented cross-linguistically (Dryer 2013; Greenberg 1963). Results will contribute to understanding how language models represent word order typological complexity attested in natural languages and how this may affect linguistic diversity globally given the widespread use of language models.

References

Anderson, Barrett R et al. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In Proceedings of the 16th Conference on Creativity & Cognition. C&C ’24, 413–425. https://doi.org/10.1145/3635636.3656204.

Booth, T.L. & R.A. Thompson 1973. Applying Probability Measures to Abstract Languages. IEEE Transactions on Computers C-22(5), 442–450. https://doi.org/10.1109/T-C.1973.223746.

Chomsky, Noam 1981. Lectures on Government and Binding: The Pisa Lectures. Berlin: Walter de Gruyter.

Dryer, Matthew S. 2013. Order of subject, object and verb. In The world atlas of language structures online. Ed. by Matthew S. Dryer & Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Universals of language. Ed. by Joseph H. Greenberg. Cambridge, MA: MIT Press, 73–113.

Guo, Yanzhu, Simone Conia, et al. 2025. Do Large Language Models have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

Papers). Ed. by Wanxiang Che et al., 3823–3838. https://doi.org/10.18653/v1/2025.acl-long.193.

Guo, Yanzhu, Guokan Shang, et al. 2024. Benchmarking Linguistic Diversity of Large Language Models. arXiv: 2412.10271 [cs.CL]. url: https://arxiv.org/abs/2412.10271. Vaswani, Ashish et al. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, 5998–6008.

White, Jennifer & Ryan Cotterell 2021. Examining the inductive bias of neural language models with artificial languages. In Proceedings of ACL-IJCNLP, 454–463.

Linguistic diversity and digitalization: a progress report

Juliane Benson, Julia Neidhardt, Katharina Zeh, Andreas Baumann, Hannes Essfores, Hannes Fellner

Linguistic diversity is crucial for maintaining cultural heritage and human knowledge, yet languages worldwide are rapidly disappearing (Simons, 2019). Previous research has mostly focused on social, economic, and environmental drivers of this decline (Bromham et al., 2021). While it is plausible that digitalization might have an impact on linguistic diversity (Cunliffe, 2017), the directionality of this impact remains unclear.

In this presentation we show the newest advances of our project Disentangling effects of digitalization on linguistic diversity. We address different approaches and research questions to explore the relationship between digitalization and linguistic diversity (Benson et al., forthcoming). In line with our focus on linguistic diversity, we also examine the effects on language endangerment. A key part of the project is collecting and processing data, including gathering digitalization indices (Zeh, 2025) and web scraping linguistic data from *Ethnologue* (Benson, forthcoming). These data serve as the basis for our analyses, which include analyzing the relationship between global measures of linguistic diversity and digitalization indices at the country level (Zeh, 2025), comparing linguistic diversity in digital and non-digital spheres, and applying a digital approach to examine linguistic diversity. Additionally, we conduct a diachronic case study with a focus on Canada (Benson Forthcoming), examining non-digital linguistic diversity over the last 30 years and collecting recent interviews on personal experiences of linguistic diversity in the digital space.

References

Benson, Juliane, Katharina Zeh, Hannes Essfors, Hannes Fellner, Julia Neidhardt & Andreas Baumann. Forthcoming. Linguistic diversity and digitalization: an ambivalent relationship. (Paper to be presented at the Digital Humanism Interdisciplinary Science and Research Conference 2025 (DIGHUM-RES), Vienna, 20–21 November 2025.)

Benson, Juliane. Forthcoming. *Diachronic Linguistic Diversity in Canada. Vienna, Austria: University of Vienna MA thesis.*

Bromham, Lindell, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill & Xia Hua. 2021. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution* 6(2). 163–173. https://doi.org/10.1038/s41559-021-01604-y

Cunliffe, Daniel. 2007. Minority languages and the Internet: new threats, new opportunities. In Mike Cormack & Niamh Hourigan (eds.), *Minority language media*, 133–150. Clevedon: Multilingual Matters. https://doi.org/10.21832/9781853599651-008

Simons, Gary F. 2019. Two centuries of spreading language loss. *Proceedings of the Linguistic Society of America* 4(1). 27. https://doi.org/10.3765/plsa.v4i1.4532

Zeh, Katharina. 2025. *Quantifying the Relationship Between Digitization and Linguistic Diversity: A Multilevel Statistical Analysis Using R.* Vienna, Austria: University of Vienna MA thesis.

The rise of present participial -ing in Middle English: Reducing ambiguity in word structure Signals

Michelle van de Bilt, Florian Jung, Matthis Hupertz, Irene Böhm, Nikolaus Ritt

We explore the role of sound patterns as signals of word structure in language change, using digital databases and AI-supported statistical analyses.

When Middle English underwent system-wide schwa loss, this disrupted how phonotactic patterns served as cues to word and morpheme boundaries. One consequence was that the word-final /-nd/ cluster became an increasingly ambiguous signal: It could indicate a word boundary (*find*), a morpheme boundary in past tense forms (*burn+ed*), or a present participial suffix (*barn+inde* 'burning'). As such ambiguity hinders word processing (Post et al. 2008) and is thus dispreferred in language change (Dressler et al. 2010, Baumann et al. 2019, Böhm et al. accepted), we propose that the debated (Budna 2014) shift from the Old English present participle *-ende* to Modern English *-ing* may have been partly motivated by pressures to reduce it.

To investigate this, we used the *ECCE* database (Ritt et al. 2018), an interactive digital resource based on the Penn-Helsinki Parsed Corpora of Middle English (Kroch & Taylor 2000) and Early Modern English (Kroch et al. 2004). We extracted type counts for *-nd* wordforms and *-ing* present participles from 1150 to 1699, quantified the ambiguity of *-nd* via Shannonian entropy, and related it to the spread of present participial *-ing* through time-series, Spearman's rank correlation, and Granger causality analyses with the assistance of generative AI (OpenAI 2024).

Our findings tentatively suggest a connection between the ambiguity of *-nd* and the rise of present participial *-ing*: As *-ing* spread, the ambiguity of *-nd* decreased. Granger causality tests indicated a significant predictive relationship from *-ing* frequency to *-nd* ambiguity; however, the reverse relationship was not substantiated, providing only limited support for our hypothesis. In our talk, we will discuss these insights in greater detail and show how our research was made possible by existent digital resources and AI.

References

Baumann, Andreas, Christina Prömer & Nikolaus Ritt. 2019. Word form shapes are selected to be morphonotactically indicative. *Folia Linguistica Historica* 40(1). 27–52. https://doi.org/10.1515/flih-2019-0007

Böhm, Irene, Nikolaus Ritt & Theresa Matzinger. accepted. Sound changes are selected by a bias against morphotactic ambiguity. *Journal of Historical Linguistics*.

Budna, Anna. 2014. The present participle mark-ing in East Midland Middle English: a corpus study. *International Journal of English Studies* 23(2). 42–51.

Dressler, Wolfgang U., Katarzyna Dziubalska-Kołaczyk & Lina Pestal. 2010. Change and variation in morphonotactics. *Folia Linguistica Historica* 44(31). 51–67. https://doi.org/10.1515/flih.2010.003

Kroch, Anthony & Ann Taylor. 2000. *Penn-Helsinki Parsed Corpus of Middle English* [Corpus]. University of Pennsylvania. https://ling.upenn.edu/hist-corpora/

Kroch, Anthony, Beatrice Santorini & Lauren Delfs. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English* [Corpus]. University of Pennsylvania. https://www.ling.upenn.edu/hist-corpora/

OpenAI. 2024. *ChatGPT,* GPT-4, version June 2025 [Large language model]. https://chat.openai.com/chat

Post, Brechtje, William D. Marslen-Wilson, Billi Randall & Lorraine K. Tyler. 2008. The processing of English regular inflections: Phonological cues to morphological structure. *Cognition* 109. 1–17. https://doi.org/10.1016/j.cognition.2008.06.011

Ritt, Nikolaus, Andreas Baumann & Christina Prömer. 2018. *ECCE: Evolution of consonant clusters in English: An interactive database of English word-final consonant clusters* [Database]. University of Vienna. https://ecce.univie.ac.at

The reliability of detecting and exploring basic emotions in short social media texts using the BEMDI-metre

Natalia Borza

To explore basic emotions (BEM) in discourse, the BEMDI-metre has been developed (Borza, forthcoming). The analytical tool is the operationalization of Ekman's (1992) psychological taxonomy of BEMs in discourse analysis. The Ekmanian taxonomy of the clusters of the seven BEMs – anger, fear, enjoyment, sadness, disgust, contempt, and surprise – allows for the exploration of 48 emotions. The discourse analytical tool, which is capable of exploring multiple emotions in texts, is context dependent, and largely language independent. The BEMDI-metre can identify both explicitly stated BEMs (detection) and BEMs that are not named in the discourse (exploration). In order to provide insight into the operational mechanisms of the analytical tool, the exploration of BEMs in Austrian social media (Facebook) samples from the 2024 European Parliament elections is presented (posts by political parties e.g. FPÖ, Grünen, Neos and their supporters' comments).

In the Second Austrian Meeting on Digital Linguistics (ÖTDL, Innsbruck, 2024) the reliability of the BEMDI-metre was demonstrated by reporting the inter-rater reliability (IRR) of six annotators as the percentage agreement. The present study further enhances knowledge on the reliability of the discourse analytical tool by analysing the same corpus of social media comments (N=45) and the annotation of the same six raters (for further details see ÖTDL, Innsbruck, 2024, and Borza, 2023) using statistics that takes expected agreement into account. The chosen measure of IRR was the Fleiss Kappa, which factors out random guessing (Everitt and Skrondal, 2010: 9). For the calculation, the online statistical calculator DATAtab (2025), developed by the Technical University of Graz, was applied. The Fleiss Kappa values, in conjunction with the upper and lower confidence intervals (CIs), were mapped onto the Landis and Koch (1977) scale.

The results of the study show that the application of the BEMDI-metre yields consistent results across all BEMs, exhibiting substantial inter-rater agreement. The consistency can reach almost perfect agreement, and attains slight agreement at the minimum level.

References

Borza, Natalia. 2023. "We have the right to choose who to live with": Discursive legitimation strategies in the Facebook comments of *Fidesz* supporters. In Koller Veronika, Borza Natalia, Demata Massimiliano, Filardo-Llamas Laura, Gustafsson Anna, Kopf Susanne, Miglbauer Marlene, Reggi Valeria, Šarić Ljiljana, Brylla Charlotta and Stopfner Maria (eds.), *Voices of Supporters: Populist Parties, Social Media and the 2019 European Elections*, 78-110. Amsterdam: John Benjamins Publishing Company.

DATAtab Team. 2025. DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria. URL https://datatab.net

Ekman, Paul. 1992. An argument for basic emotions. *Cognition and Emotion* 6(3-4): 169-200.

Everitt, Brian S. & Skrondal, Anders. 2010. *The Cambridge dictionary of statistics.* Cambridge: CUP.

Landis Richard J. & Koch Gary G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.

ÖTDL (2. Österreichisches Treffen zur Digitalen Linguistik – Second Austrian Meeting on Digital Linguistics), Innsbruck, 2024; workshop in the framework of 48. Österreichische Linguistiktagung (ÖLT) https://www.uibk.ac.at/de/congress/oelt2024/programm/

Regional Semantic Change and Variation in Ukrainian with LLMs

Nataliia Cheilytko

This presentation explores how large language models (LLMs) can be used to detect regional semantic variation and ongoing semantic change in Ukrainian, focusing on polysemous lexemes. We present a classification-based word-sense disambiguation (WSD) task applied to 200 lexemes selected from the General Regionally Annotated Corpus of Ukrainian (GRAC, www.uacorpus.org, Shvedova et al. 2017-2025), a curated panchronic reference corpus of Ukrainian with complex regional annotation (Shvedova & von Waldenfels 2021) specifically designed to facilitate research into regional variation of written Ukrainian.

These lexemes, annotated with predefined sense inventories, were processed in context using GPT-4o to classify each occurrence according to its appropriate sense, as proposed in Cheilytko & von Waldenfels 2024b.

The analysis revealed a notable subset of "non-classified" occurrences — instances in which the LLM failed to assign any listed sense from the inventory. These cases, far from being random errors, emerged as linguistically meaningful data points. Subsequent manual and computational linguistic analysis of these outliers revealed several patterns of regional variation and semantic innovation. In particular, certain polysemous lexemes displayed shifts in dominant senses depending on geographical context (which confirms initial observations made in Cheilytko & von Waldenfels 2024a), while others showed clear evidence of semantic extension or drift not yet captured in existing lexicographical resources.

By leveraging the interpretive capabilities of LLMs not only to classify but also to highlight non-conforming usage, this study demonstrates a novel method for tracing semantic change and regional differentiation in under-resourced languages. The project also raises questions about the dynamic interplay between language models and traditional semantic annotation, and how AI tools might reshape linguistic fieldwork in digital corpora.

References

Cheilytko & von Waldenfels 2024a — Cheilytko N., von Waldenfels R. (2024). Word Embeddings for Detecting Lexical Semantic Change in Ukrainian. Lexicography and Semantics Proceedings of the XXI EURALEX International Congress 8–12 October 2024 Cavtat, Croatia, pp. 231-241. Available at https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralax-XXI-final-web.pdf

Cheilytko & von Waldenfels 2024b — Cheilytko N., von Waldenfels R. (2024). Semantic Change and Lexical Variation in Ukrainian with Vector Representations and LLM. Book of Abstracts of the Workshop Large Language Models and Lexicography. Simon Krek (Ed.), pp. 1-5. Available at https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf

Shvedova et al. 2017-2025 — Heneral'nyj rehional'no anotovanyj korpus ukraïns'koï movy (HRAK) / M. Švedova, R. fon Val'denfel's, S. Jaryhin, A. Rysin, V. Starko, T. Nikolajenko, A. Lukaševs'kyj ta in. — Kyïv, L'viv, Jena, 2017–2025. — uacorpus.org.

Shvedova & von Waldenfels 2021 — Shvedova [Švedova] M., von Waldenfels R. Regional Annotation within GRAC, a Large Reference Corpus of Ukrainian: Issues and Challenges. In: *CEUR Workshop Proceedings. Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*. Volume I: Main Conference. Kharkiv, Ukraine, April 22-23, 2021. pp. 32-45. Available at http://ceur-ws.org/Vol-2870/paper4.pdf

Phylogenetic analysis of folktale evolution: Reconstructing cultural transmission through computational Methods

Edlira Gugu

This study applies phylogenetic analysis, a method developed for biological evolution and later adapted for historical linguistics, to analyze the co-evolution and transmission of linguistic and literary traditions. The primary objective is to demonstrate how computational methods can identify historical relationships among myths, folktales, and their linguistic carriers across different cultures, exploring the parallel transmission patterns of language families and narrative traditions.

The research integrates linguistic phylogeny with literary analysis through two main corpora: (1) 175 versions of the "Cosmic Hunt" myth, analyzed using phylogenetic networks and Bayesian reconstruction, and correlated with Indo-European and other language family trees; (2) 58 versions of "Little Red Riding Hood" (ATU 333) from 33 linguistically diverse cultures, evaluated through 72 narrative variables alongside linguistic distance measurements using cladistic, Bayesian and phylogenetic network-based methods.

Phylogenetic analysis reveals strong correlations between narrative evolution and linguistic phylogenies. The "Cosmic Hunt" myth's distribution correlates significantly with Indo-European language dispersals, evidencing Paleolithic co-transmission of linguistic and mythological elements along human migration routes. The "Little Red Riding Hood" study demonstrates that folktale transmission follows linguistic family boundaries more closely than geographic proximity, with tale variants clustering according to language family relationships rather than spatial distribution. African tales classify as variants of "The Wolf and the Kids" (ATU 123), corresponding to Afro-Asiatic linguistic patterns, while East Asian versions show narrative hybridization paralleling language contact zones.

This interdisciplinary approach demonstrates that linguistic and literary traditions co-evolve through shared transmission mechanisms. Phylogenetic methods reveal how narrative structures adapt to linguistic constraints while maintaining core semantic content. The correlation between language family trees and folktale phylogenies provides evidence for vertical cultural transmission alongside linguistic inheritance, offering new insights into the relationship between cognitive-linguistic structures and narrative universals. These studies demonstrate that narratives evolve through processes similar to biological and linguistic ones, including punctuated equilibrium and a strong correlation between geographic distribution and phylogenetic distance. This methodology opens new perspectives for cultural transmission studies and language-literature relationships.

References

Bortolini, Eugenio, Luca Pagani, Enrico R. Crema, Sara Sarno, Chiara Barbieri, Alberto Boattini, et al. 2017. Inferring patterns of folktale diffusion using genomic data. *Proceedings of the National Academy of Sciences* 114(34).9140–9145.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexei V. Alekseyenko, Andrew J. Drummond, et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097).957–960.

Campbell, Joseph. 1949. *The Hero with a Thousand Faces*. Princeton, NJ: Princeton University Press.

d'Huy, Julien. 2013a. A phylogenetic approach of mythology and its archaeological consequences. *Rock Art Research* 30(1).115–118.

d'Huy, Julien. 2013b. A Cosmic Hunt in the Berber sky: a phylogenetic reconstruction of Palaeolithic mythology. *Les Cahiers de l'AARS* 15.93–106.

da Silva, Sara G., & Jamshid J. Tehrani. 2016. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science* 3(1).150645.

Gray, Russell D., & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965).435–439.

Greenhill, Simon J., Chih-Han Wu, Xuhua Hua, Michael Dunn, Stephen C. Levinson, & Russell D. Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114(42).E8822–E8829.

Lévi-Strauss, Claude. 1955. The structural study of myth. *Journal of American Folklore* 68(270).428–444.

Pagel, Mark. 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics* 10(6).405–415.

Propp, Vladimir. 1968. *Morphology of the Folktale*. Austin, TX: University of Texas Press. (Original work published 1928).

Ross, Rachel M., Simon J. Greenhill, & Quentin D. Atkinson. 2013. Population structure and cultural geography of a folktale in Europe. *Proceedings of the Royal Society B* 280(1756).20123065.

Tehrani, Jamshid J. 2013. The phylogeny of Little Red Riding Hood. *PLOS ONE* 8(11). e78871.

Thompson, Stith. 1955–1958. *Motif-Index of Folk-Literature* (6 volumes). Bloomington, IN: Indiana University Press.

Thuillard, Morgan, Jean-Loïc Le Quellec, & Julien d'Huy. 2018. Computational approaches to myths analysis: Application to the Cosmic Hunt. *Nouvelle Mythologie comparée* 4.1–32.

Witzel, E. J. M. 2012. *The Origins of the World's Mythologies*. Oxford: Oxford University Press.

Potential of Generative AI for Text Transcription

Jona Marie Hassenbach, Magdalena Miteva

Recent advances in generative AI have prompted interest in their possible applications in the field of digital linguistics. As part of the PressMint project, we compared the performance of various large language models (LLMs) for the transcription of historical newspapers. As a gold standard, we used texts initially transcribed by Transkribus and subsequently corrected by humans. Against this gold standard, we evaluated texts transcribed by OpenAI, Google Vision, Gemini, and Anthropic, using a zero-shot procedure and testing different prompts. We also included the initial texts produced by Transkribus, as well as an earlier transcription provided by Anno, the database of the Austrian National Library, in our comparison.

For evaluation, we employed standard OCR metrics such as Character Error Rate (CER) and Word Error Rate (WER), both of which are derived from the Levenshtein distance (Levenshtein 1966), and the Ratcliff/Obershelp pattern recognition algorithm (Ratcliff & Metzener 1988). This was supplemented by qualitative error analysis of typical challenges in Fraktur script.

The results show that specialized OCR tools like Transkribus still outperform LLMs. In two out of three metrics (CER and WER), Anno and Transkribus achieved similarly high results, while

Gemini, the highest-performing LLM, performed significantly worse. Interestingly, Gemini ranked second on the third metric, the pattern recognition algorithm, where Anno yielded much poorer results. This indicates that, while similar to Anno in terms of character positioning, Transkribus (and to some extent Gemini) is substantially better at correctly recognizing longer sequences of text. Overall, our findings suggest that while LLMs cannot yet replace dedicated OCR systems, they hold potential as complementary tools, particularly for post-correction, multilingual processing, or hybrid transcription workflows in the field of digital humanities.

References

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8). 707-710.

Ratcliff, John W. & Metzener, David E. 1988. Pattern Matching: The Gestalt Approach. Dr. Dobb's Journal 13(7). 46-72.

Models

Anthropic. 2025. Claude Sonnet 4. https://www.anthropic.com/news/claude-4 (2 October 2025)

Google Cloud. 2025. Cloud Vision API. https://cloud.google.com/vision/docs?hl=de (2 October 2025)

Google DeepMind. 2025. Gemini 2.5 Flash-Lite. https://deepmind.google/models/gemini/flash-lite/ (2 October 2025)

OpenAI. 2024. GPT-4o. https://openai.com/de-DE/index/hello-gpt-4o/ (2 October 2025)

Vergleichende Analyse von nominierten Texten des Ingeborg-Bachmann-Preises

Katharina Horn, Fabian Navarro, Jasmin Bettstein

In unserem Vortrag gehen wir der Frage nach, ob sich Korrelationen zwischen den Komplexitätsmaßen eines literarischen Textes und einer Auszeichnung im Rahmen der *Tage der deutschsprachigen Literatur* (kurz: TDDL), auch bekannt als *Ingeborg-Bachmann-Preis*, feststellen lassen. Zu diesem Zweck haben wir ein Korpus erstellt, das alle digital verfügbaren Texte der TTDL von 1999 bis 2025 enthält. Diese Fragestellung knüpft an die Arbeit von Karin Röhricht aus dem Jahr 2016 an. Röhricht fokussiert sich in ihrer Arbeit jedoch auf eine Inhalts- und Themenanalyse und beschränkt sich lediglich auf jene Texte von 1977 bis 2011, die im Rahmen der jährlich erscheinenden Anthologie *Klagenfurter Texte* veröffentlicht wurden. Unser Datensatz wurde hingegen vor dem Hintergrund einer sprachwissenschaftlichen Analyse erstellt und beinhaltet alle digital verfügbaren Texte, die auf der offiziellen Seite der TTDL veröffentlicht wurden.

Mithilfe eines R-Skripts berechnen wir den *Flesch-Grad,* den *Hapax legomenon,* die lexikalische Diversität (*Type-Token Ratio*, TTR), die lexikalische- und grammatische Dichte sowie das Lemma-Token-Verhältnis (*Lemma-Token Ratio*, LTR). Unsere vorläufigen Ergebnisse zeigen keine signifikanten Unterschiede zwischen ausgezeichneten und nicht ausgezeichneten Texten. Dies könnte nahelegen, dass bei der Bewertung von literarischen Texten die genannten Komplexitätsmaßen wenig bis keine Rolle spielen. Im Zuge unseres Projekts ist eine Datenbank entstanden, die für die Jahre von 1999 bis 2025 Informationen wie Name, Land, Titel, Preis, Preisgeld und Dateiformat der nominierten Texte enthält. Sie ist somit um kommende Ausgaben der TDDL erweiterbar und stellt eine vielversprechende Datenbasis für künftige quantitative und qualitative Analysen dar.

Referenzen

Flesch, Rudolf. 1948. *A New Readability Yardstick*. Journal of Applied Psychology. 32, Nr. 3, 221–233. ORF. 2025. Bachmann-Preis-Archiv von 1997 bis heute. https://bachmannpreis.orf.at/stories/archiv/ (04. September, 2025.)

Röhricht, Karin. 2016. *Wettlesen um den Ingeborg-Bachmann-Preis: Korpusanalyse der Anthologie "Klagenfurter Texte" (1977-2011)*. Innsbruck: Studien Verlag.

Kommunikationsverben als Indikatoren für situativen Kontext – Ein Forschungsprojekt zu Indexikalität in Korpusanalyse und Sprachwissen

Cordula Meißner, Janina Deilke, Anna-Lena Randermann

Der Vortrag stellt das vom FWF geförderte Forschungsprojekt *Kommunikationsverben als Indikatoren für situativen Kontext* vor. Dieses zielt darauf ab, die Annahme gebrauchsbasierter Sprachmodelle und ihrer methodologischen Einlösung durch die korpuslinguistische Analyse anhand des Phänomens der Indexikalität zu überprüfen. Indexikalität fasst das Phänomen, dass die Verwendung eines sprachlichen Ausdrucks bestimmte situative Kontexte signalisieren oder evozieren kann. Aus der Perspektive gebrauchsbasierter Sprachmodelle ergibt sich die Erwartung, dass sich die Indexikalität

sprachlicher Ausdrücke aus ihrem wiederholten Auftreten in bestimmten situativen Kontexten entwickelt, d. h., dass Sprecher:innen Gemeinsamkeiten aus wahrnehmbaren Ko-Vorkommen sprachlicher und situativer Merkmale abstrahieren und auf dieser Grundlage ihr sprachliches Wissen über Indexikalität aufbauen (vgl. Schmid, 2020). Die korpuslinguistische Analyse wird als methodologische Umsetzung der Annahmen gebrauchsbasierter Modelle angesehen, da sie die Untersuchung sprachlicher Phänomene hinsichtlich ihrer Häufigkeit im natürlichen Sprachgebrauch ermöglicht (vgl. Stefanowitsch, 2011) und im Sinne der Korpus-als-Input-Perspektive Rückschlüsse auf das Wissen der Sprachgemeinschaft zulassen könnte (vgl. Stefanowitsch & Flach, 2017). Im Projekt soll anhand des Wortschatzes der Kommunikationsverben (Harras et al. 2004) aufbauend auf einer korpuslinguistischen Pilotstudie (Meißner 2025) mittels der Triangulation von Korpusanalyse und experimenteller Erhebung von Sprachwissen untersucht werden, inwieweit diese Annahme zutrifft.

Referenzen

Harras, Gisela, Edeltraud Winkler, Sabine Erb & Kristel Proost. 2004. *Handbuch deutscher Kommunikationsverben:Teil 1: Wörterbuch* (Schriften des Instituts für Deutsche Sprache10.1). Berlin, New York: Walter de Gruyter.

Meißner, Cordula. 2025. Muster domänenbezogener Indexikalität von Kommunikationsverben im gesprochenen

Deutsch: Ein korpuslinguistischer Beschreibungsansatz. In Nadine Proske, Tilo Weber, Arnulf Deppermann & Monika Dannerer (eds.), *Gesprochenes Deutsch: Struktur, Variation, Interaktion* (Jahrbuch des Instituts für Deutsche Sprache 2024), 241–266. De Gruyter.

Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment.* Oxford, New York: Oxford University Press.

Stefanowitsch, Anatol. 2011. Cognitivelinguistics meets the corpus. In Mario Brdar, Stefan T. Gries & Milena Žic Fuchs (eds.), *Cognitive linguistics: Convergence and expansion* (Human Cognitive Processing 32), vol. 32, 257–290. Amsterdam: Benjamins.

Stefanowitsch, Anatol & Susanne Flach. 2017. The corpus-based perspective on entrenchmen. In Hans J. Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (Language and the human lifespan series), 101–128. Berlin, Boston, Washington, DC: De Gruyter Mouton; American Psychological Association.

Philosophical considerations of digital text analysis: Core assumptions and methodological challenges

Rashid Mustafin

The main advantage of digital text analysis and its raison d'être is the possibility to process large sets of texts automatically. The nature of digital text analysis and the data that it is most typically used on, textual digital trace data, imply certain assumptions and methodological challenges. This report reflects on these assumptions and challenges through the critical examination of digital text analysis in terms of philosophy of science. The arguments will be illustrated by a few “cautionary tale” examples based on the existing critiques (Andreski 1972, Baden et al. 2022, Da 2019, Loughran & McDonald 2011) and personal experience in reading and conducting research. The underlying assumptions to be discussed include the capacity of the textual data to represent reality in both fundamental and practical senses, the appropriateness of using measurable proxies for making conclusions about abstract phenomena, and the value of gaining “a big picture” through distant reading. Apart from that, the report will also overview some of the current challenges in digital text analysis; namely, the problem of authenticity and representativeness of digital trace data and the intransparent nature of deep learning models that gain popularity in contemporary research.

References

Andreski, Stanislav. 1972. Social Sciences as Sorcery. London: André Deutsch Limited.

Baden, Chirsitan, Christian Pipal, Martijn Schoonvelde, & Mariken A. C. G. van der Velden. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. Communication Methods and Measures 16(1), 1-18.

Da, Nan Z. 2019. The Computational Case against Computational Literary Studies. Critical Inquiry 45, 601-639.

Loughran, Tim & Bill McDonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66, 35-65.

Large language models as synthetic participants in psycholinguistic experiments: the case of German noun plural formation

Teodor Petrič

The present study investigates whether large language models (LLMs) can provide synthetic psycholinguistic data that approximate the responses of human learners. We focus on the domain of German plural formation, a classical testbed in psycholinguistics and morphology (McCurdy et al. 2020; Dankers et al. 2021; Sauerland et al. 2025). Using the 24 nonce-noun stimuli of Marcus et al. (1995), extended across all three grammatical genders (72 items total), we collected responses from 123 Slovenian learners of German as a second language (L2) and compared them with outputs from several locally run LLMs (via Ollama and LM Studio). Each model was prompted to supply genitive and plural forms in context; for every stimulus, at least ten “synthetic participants” were generated per model.

Responses were coded for plural class (-(e)n, -e, -er, -s, zero, other) and umlaut application. We computed distributional similarity between students and LLMs using Jensen–Shannon divergence, tested equivalence of class rates via two one-sided tests (TOST), and fitted mixed-effects logistic regression models for umlaut and minority classes (-s, -er).

Preliminary analyses (including also other plural formation studies: e.g., Zaretsky et al. 2013, 2016; Schuhmann & Smith 2024) suggest that some LLMs reproduce the *overall class distribution* of human L2 learners reasonably well, especially in overgeneralization patterns (preference for -(e)n and avoidance of minority classes). However, divergences remain: models differ in their sensitivity to gender cues and in their variability across items. Our findings highlight both the *promise and limitations* of using LLMs (Wilcox et al. 2025; Machowald et al. 2024) as stand-ins for human participants in psycholinguistic experiments. We argue that under controlled conditions, LLM outputs may complement (but not replace) human data, serving as a cost-effective means for pretesting and hypothesis exploration.

References

Dankers, Verna & Langedijk, Anna & McCurdy, Kate & Williams, Adina & Hupkes, Dieuwke. 2021. Generalising to German Plural Noun Classes, from the Perspective of a Recurrent Neural Network. Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL). 94–108.

Marcus, Gary & Brinkmann, Ursula & Clahsen, Harald & Wiese, Richard & Pinker, Steven. 1995. German inflection: The exception that proves the rule. Cognitive Psychology 29(3). 189–256.

Mahowald, Kyle & Ivanova, Anna & Blank, Idan & Kanwisher, Nancy & Tenenbaum, Joshua & Fedorenko, Evelina. 2024. Dissociating language and thought in large language models. arXiv:2301.06627v3.

McCurdy, Kate & Goldwater, Sharon & Lopez, Adam. 2020. Inflecting when there’s no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. arXiv:2005.08826v1.

Petrič, Teodor. 2025. Pluralmarkierung im Deutschen als Fremdsprache. In Simoska, Silvana (ed.), Grundzüge deutscher Sprachstrukturen in Wort, Satz und Text. Festschrift zu

Ehren des 90. Geburtstages von Prof. Dr. Wolfgang Motsch. Skopje: Hll.-Kyrill-und-Method-

Universität. 243 - 256. https://flf.ukim.mk/wp-content/uploads/Festschrift-Endversion-03.09-5.2.4-FINAL.pdf.

Sauerland, Uli & Matthaei, Celia & Salfner, Felix. 2025. Child vs. machine language learning: Can the logical structure of human language unleash LLMs? arXiv:2502.17304v1.

Schuhmann, Katharina & Smith, Laura. 2024. From formalism to intuition: probing the role of the trochee in German nominal plural forms in L1 and L2 German speakers. Frontiers in Language Sciences Volume 3. https://doi.org/10.3389/flang.2024.1338625.

Wilcox, Ethan & Hu, Michael & Mueller, Aaron & Warstadt, Alex & Choshen, Leshem & Zhuang, Chengxu & Williams, Adina & Cotterell, Ryan & Linzen, Tal. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. Journal of Memory and Language 144. https://doi.org/10.1016/j.jml.2025.104650.

Zaretsky, Eugen & Neumann, Katrin & Euler, Harald & LANGE, BENJAMIN. 2013. Pluralerwerb im Deutschen bei russisch- und türkischsprachigen Kindern im Vergleich mit anderen Migranten und monolingualen Muttersprachlern. Zeitschrift für Slawistik 58(1). 43–71.

Zaretsky, Eugen & Müller, Hans-Helge & Lange, Benjamin. 2016. No default plural marker in Modern High German. Italian Journal of Linguistics, 28(2). 203-230.

Processing Digitized Text on an Example of Job Advertisements from Austrian Periodicals from 1850-1950

Klara Venglerova

This doctoral thesis presents a comprehensive pipeline designed within the JobAds project for OCR and text mining of Austrian historical newspapers, with a particular focus on job advertisements published between 1850 and 1950. It starts with considerations about corpus creation, including its representativeness (Biber 1993; Reppen 2022; Atkins, Clear & Ostler 1992; Bauer & Aarts 2000) and contextualization and bias (Beelen et al. 2022), and continues with the process of turning text within images into machine-readable text. This typically covers layout analysis, OCR, and post-correction.

For the layout analysis, a comprehensive evaluation framework has been developed (Venglarova et al. 2024) and used to evaluate several models, from which Eynollah (Rezanezhad et al. 2023) yields the best results. Also, several OCR models have been evaluated, including different pre-processing techniques, such as binarization using Otsu threshold (Gupta, Jacobson & Garcia 2007; Chang & ZhiYing 2009). The best results on our data, which are a mixture of Fraktur and Antigua, were achieved by the frak2021 model (Mannheim University Library 2021) with the mean CER 0.155 on the individual job advertisements, for which we manually created the ground truth.

Subsequent tasks include text type classification, where each text segment is classified as a job advertisement or not, and further subdivided into, e.g., job searches and job offers categories. From job advertisements, job titles are extracted (Adam, Venglarova & Vogeler 2025) to be further used in an economic analysis. An overview of potential use-cases of the extracted data, based on an interview with experts from the field of economics, concludes the work.

References

Adam, Raven, Klara Venglarova & Georg Vogeler. 2025. Exploring Historical Labor Markets: Computational Approaches to Job Title Extraction. Journal of Data Mining & Digital Humanities NLP4DH. https://doi.org/10.46298/jdmdh.15038.

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus Design Criteria. Literary and Linguistic Computing 7. 1–16. https://doi.org/10.1093/llc/7.1.1.

Bauer, Martin W. & Bas Aarts. 2000. Corpus construction: a principle for qualitative data collection. In. https://api.semanticscholar.org/CorpusID:158558066.

Beelen, Kaspar, Jon Lawrence, Daniel C S Wilson & David Beavan. 2022. Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. Digital Scholarship in the Humanities 38(1). 1–22. https://doi.org/10.1093/llc/fqac037.

Biber, Douglas. 1993. Representativeness in corpus design. Literary and Linguistic Computing 8. 243–257.

Chang, Loh Zhi & Steven Zhou ZhiYing. 2009. Robust pre-processing techniques for OCR applications on mobile devices. In Proceedings of the 6th International Conference on Mobile Technology, Application & Systems (Mobility ’09). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1710035.1710095.

Gupta, Maya R., Nathaniel P. Jacobson & Eric K. Garcia. 2007. OCR binarization and image pre-processing for searching historical documents. Pattern Recognition 40(2). 389–397. https://doi.org/10.1016/j.patcog.2006.04.043.

Mannheim University Library. 2021. frak2021. https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/tessdata_best/frak2021-0.905.traineddata. (20 November, 2023).

Reppen, Randi. 2022. Building a corpus: what are key considerations? In The Routledge handbook of corpus linguistics, 13–20. Routledge.

Rezanezhad, Vahid, Konstantin Baierer, Mike Gerber, Kai Labusch & Clemens Neudecker. 2023. Document Layout Analysis with Deep Learning and Heuristics. In Proceedings of the 7th International Workshop on Historical Document Imaging and Processing HIP 2023, San José, US, August 25-26, 2023, ACM. https://doi.org/10.1145/3604951.3605513.

Venglarova, Klara, Raven Adam, Saranya Balasubramanian & Georg Vogeler. 2024. Quantifying Page Segmentation Quality in Historical Job Advertisements Retrieval. https://inria.hal.science/hal-04560463.

Interoperable historical newspapers: the PressMint Project

Tanja Wissik, Maciej Ogrodniczuk, Petya Osenova

This submission will report on the PressMint Project, which aims to compile multilingual, comparable, annotated, translated and interoperable corpora of European historical newspapers from around the start of the 20th century. The PressMint Project is funded by CLARIN, a European digital research infrastructure that offers data, tools and services to support research based on language resources. While historical newspapers are of interest to a diverse group of researchers from the social sciences and humanities (e.g., historians, historical linguists, social scientists, ethnologists, anthropologists, scholars from media and communication or cultural studies) and historical newspapers already exist for a number of languages and countries (Fišer et al., 2018) to a large extent these existing corpora are not interoperable which precludes methods for their comparison, as well as any translingual and transnational research. Therefore, the PressMint project aims to create interoperable corpora regarding metadata, annotations and formats.

PressMint project includes institutions from 17 countries (Austria, Bulgaria, Czechia, Finland, Greece, Hungary, Iceland, Italy, Latvia, Netherlands, Poland, Portugal, Slovenia, South Africa, Spain, United Kingdom and Ukraine).

The PressMint corpora will be FAIR (i.e., Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016) and linguistically annotated, including PoS tagging, lemmatization and topic classification etc. Furthermore, there will be options for including facsimiles, word normalization for historical language, and entity linking. The PressMint project builds on the successful technical framework of the ParlaMint project (Erjavec et al., 2025).

The corpora will be openly available for download in multiple formats as well as accessible via several online analysis tools.

References

Fišer, Darja, Lenardič, Jakob and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1320–1325, Miyazaki, Japan. European Language Resources Association (ELRA).

Erjavec, Tomaž, Kopp, Matyáš, Ljubešić, Nikola *et al.* 2025. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Lang Resources & Evaluation* **59**, 2071–2102 (2025). https://doi.org/10.1007/s10579-024-09798-w

Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, IJsbrand J., et al. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18