Large language models as first-pass filters: from small-sample validation to full-dataset implementation

Vítor Míguez
Instituto da Lingua Galega / Department of Galician Philology
University of Santiago de Compostela

Manual annotation usually represents the most time-consuming phase of corpus-based linguistic studies. While recent research demonstrates promising performance of large language models (LLMs) on small validation samples (Morin & Marttinen Larsson, 2025; Yu et al., 2024), the transition from experimental validation to production-scale implementation remains largely unexplored. This paper addresses two research questions: (1) How well do LLM performance results from small validation samples predict performance on complete datasets? (2) What are the practical implications of scaling LLM-assisted corpus annotation from experimental applications to production workflows?

Using semantic disambiguation of Galician *pobo* 'people/village' as a test case, we will automatically annotate a complete 6,293-example corpus with the LLM Claude 4 Opus (Anthropic, 2025). Our method prioritizes recall over precision in prompt engineering and model validation, reflecting the reality that false negatives are more costly than false positives in filtering tasks. Initial validation on 300 examples achieved 97.3% recall (F2 = 0.857) against human consensus.

To address the scalability question, in this paper we will conduct comprehensive human annotation of the complete dataset, enabling full validation of LLM performance at corpus scale. This approach will reveal whether small-sample metrics reliably predict full-dataset behavior, identify systematic error patterns that may be invisible in smaller samples, and establish practical frameworks for implementing LLM-assisted workflows while maintaining scientific rigor.

This research provides essential evidence for the transition from experimental LLM validation to practical corpus annotation workflows, addressing fundamental questions about reliability and scalability in AI-enhanced corpus linguistics.

References

Anthropic. (2025). *Claude* (Version 4). https://claude.ai/

Morin, C., & Marttinen Larsson, M. (2025). Large corpora and large language models: A replicable method for automating grammatical annotation. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2024-0228

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, *29*(4), 534–561. https://doi.org/10.1075/ijcl.23087.yu

Thematic Content Indexing of Very Large Corpora with Language Models: Towards Openness, Interoperability, and Reproducibility

Jennifer Ecker, Marc Kupietz, Roman Schneider, Ngoc Duyen Tanja Tu

IDS Mannheim

The Leibniz Institute for the German Language (IDS) maintains the world's largest linguistically curated collection of written German electronic corpora: the Deutsches Referenzkorpus (DeReKo). This heterogeneous corpus spans a wide range of genres and sources and is already partially annotated with thematic metadata. However, the current classification schema is outdated and lacks interoperability with widely adopted international standards such as the Dewey Decimal Classification (DDC) or the Universal Decimal Classification (UDC). One of the key challenges lies in developing robust and high-quality mappings between such taxonomies, enabling flexible use across diverse research and application contexts.

Manual annotation of training data for thematic corpus classification is labor-intensive and resource-demanding. Traditionally, large labeled datasets are required to train supervised models. Recent advances in artificial intelligence, particularly large language models (LLMs), open new avenues for automating this process with comparatively little manual effort.

We introduce an innovative research project that explores the capabilities of state-of-the-art language models for classifying DeReKo in accordance with internationally recognized taxonomies. One central research question is: To what extent can such models be effectively employed to classify large, thematically diverse corpora (> 100 million documents) based on standardized and interoperable classification systems? In addition to traditional norm data systems such as DDC and UDC, we aim at the integration of open, community-driven taxonomies, including the category structures of Wikipedia and Wikidata. The latter offers the advantage of being language-agnostic, thereby supporting broader thematic and multilingual openness.

Furthermore, we address practical requirements for reproducibility by examining whether restricted data can be made available for training. Publishing training data for language models is crucial for reproducible research, yet many corpora cannot be released in their original form due to copyright constraints. To tackle this, we present a multi-label text classification approach using Derived Text Formats (DTF), which enable the sharing of otherwise inaccessible data by systematically removing or altering information to reduce the risk of reconstructing the original text. However, it remains unclear how such controlled information loss affects downstream text classification performance, so we empirically investigate the impact of systematic alterations to the training texts.

ChatGPT as a thematic role tagger in Spanish

M.L. Yislén Barboza Hidalgo

University of Costa Rica

The integration of Artificial Intelligence (AI) into linguistic research has raised urgent questions about the role of human expertise in Natural Language Processing (NLP). Although AI-driven models and annotated corpora have accelerated the development of intelligent question-answer systems, many NLP tasks continue to bypass linguistic theory, especially in languages other than English. Semantic role labeling in Spanish remains notably underrepresented, despite its relevance for machine learning applications. This study addresses that gap by proposing a linguistically grounded model for thematic role annotation in Spanish and using it to evaluate ChatGPT's performance in a structured annotation task. The model was developed through a three-level descriptive analysis (syntactic, grammatical, and semantic) of the 200 most frequent verbs in the COVID-19 Corpus. Corpus Linguistics provided the methodological foundation, while Lexical-Functional Grammar (LFG) offered a robust theoretical framework. Crucially, the focus was not on fine-tuning ChatGPT, but on testing its raw linguistic performance against trained human annotators. The core objective was to measure the reliability of ChatGPT in a linguistically demanding task, using Fleiss' kappa and precision as evaluation metrics. Results indicate that ChatGPT performed moderately ($\kappa = 0.420$, precision = 0.539), falling short of human consistency ($\kappa = 0.600$, precision = 0.700). Given these results, can AI replace and potentially outperform traditional corpus-linguistic software such as AntConc or WordSmith? While AI models offer flexibility and automation, they lack the analytical transparency and linguistic precision of dedicated tools used by experts in language, making them best suited as complements. How reliably can AI perform labor-intensive annotation tasks? This study shows that although AI can assist greatly in such work, its performance still requires human oversight when high linguistic accuracy is expected for academic research.

References

Aldeen, M., Luo, J., Lian, A., Zheng, V., Hong, A. y Yetukuri, P. (2023). "ChatGPT vs. Human Annotators: A Comprehensive Analysis of ChatGPT for Text Annotation," *In International Conference on Machine Learning and Applications (ICMLA)*, pp. 602-609. doi: https://ieeexplore.ieee.org/document/10460013

Awan, A. A. (2024, July 29). *What is GPT-4 and why does it matter?*. DataCamp. Recuperado el 12 de febrero de 2025 de https://www.datacamp.com/blog/what-we-know-gpt4

Barboza Hidalgo, G.Y. (2023). *Aproximación lingüística en el diseño de un corpus anotado en español sobre COVID-19 para sistemas de pregunta-respuesta*. [Tesis de maestría, Universidad de Costa Rica]. https://www.kerwa.ucr.ac.cr/items/64251e04-9fbb-4721-8d33-4eb01ec83f97

Benites, L. (2021, 17 de noviembre). *Muestra Bootstrap: Definición, ejemplo*. Statologos. Recuperado el 4 de junio de 2023, de https://statologos.com/muestra-de-arranque/

Benites, L. (2022, 7 de junio). *Estimación de confiabilidad Alfa de Krippendorff: Definición simple*. Statologos. Recuperado el 04 de junio de 2023, de https://statologos.com/alfa-de-krippendorff/

Bresnan, J., Asudeh, A., Toivonen, I. y Wechsler, S. (2016). *Lexical-Functional Syntax*. (2da ed.). Blackwell Publishers Ltd., West Sussex, UK.

Castellón, I., Climent, S., Coll-Florit, M., Lloberes, M. y Rigau, G. (2012). Constitución de un corpus de semántica verbal del español: Metodología de anotación de núcleos argumentales. *RLA. Revista de lingüística teórica y aplicada, 50* (1), 13-38. https://dx.doi.org/10.4067/S0718-48832012000100002

Chan, C., Cheng, J., Wang, W., Jiang, Y., Fang, T., Liu, X. y Song, Y. (2023). Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*. https://doi.org/10.48550/arXiv.2304.14827

Colic, N. y Rinaldi, F. (2019). Improving spaCy dependency annotation and PoS tagging web service using independent NER services. *Genomics & Informatics, 17*(2). Recuperado el 15 de junio de 2022, de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6808626/

Dilmegani, C. (6 de enero de 2025). *How to build a chatbot: Components & Architecture in 2025*. AIMultiple. Recuperado el 11 de febrero de 2025 de https://research.aimultiple.com/chatbot-architecture/

Fichtel, L., Kalo, J. C., y Balke, W. T. (2021). Prompt tuning or fine-tuning-investigating relational knowledge in pre-trained language models. In *3rd Conference on Automated Knowledge Base Construction*. Recuperado el 24 de mayo de 2023, de https://openreview.net/forum?id=o7sMlpr9yBW

García Marchena, O. (2009). Lingüística española e Inteligencia Artificial Aplicación informática de gramáticas de restricciones para la confección de agentes de diálogo. *Interlingüistica, (18)*, 472-483. Recuperado el 15 de junio de 2022, de https://dialnet.unirioja.es/servlet/articulo?codigo=3130252

George, A. S., George, A. H. y Martin, A. G. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal*, *1*(1), 9-23. https://doi.org/10.5281/zenodo.7644359

Gerdes, K. y Kahane, S. (2016). Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 131-140). https://halshs.archives-ouvertes.fr/halshs-01509118/

Gilardi, F., Alizadeh, M. y Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*. https://doi.org/10.48550/arXiv.2303.15056

Gil-Vallejo, L., Castellón, I., Coll-Florit, M. y Turmo, J. (2015). Hacia una clasificación verbal automática para el español: estudio sobre la relevancia de los diferentes tipos y configuraciones de información sintáctico-semántica. *Linguamática, 7*(1), 41-52. Recuperado el 15 de junio de 2022, de https://upcommons.upc.edu/handle/2117/77807

Hariri, W. (2023). Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. *arXiv preprint arXiv:2304.02017*. https://doi.org/10.48550/arXiv.2304.02017

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... y Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*. https://doi.org/10.48550/arXiv.2302.09210

Huang, F., Kwak, H. y An, J. (2023). Is chatgpt better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*. https://doi.org/10.48550/arXiv.2302.07736

Johnson. (26 de abril de 2023). *CHATGPT raises questions about how humans acquire language*. The Economist. Recuperado el 21 de mayo de 2023, de https://www.economist.com/culture/2023/04/26/chatgpt-raises-questions-about-how-humans-acquire-language

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., ... y Kazienko, P. (2023). Chatgpt: Jack of all trades, master of none. *Information Fusion, 99.* https://doi.org/10.1016/j.inffus.2023.101861

Koubaa, A., Boulila, W., Ghouti, L., Alzahem, A. y Latif, S. (2023). Exploring ChatGPT Capabilities and Limitations: A Critical Review of the NLP Game Changer. *Preprints.org*, 2023030438. https://doi.org/10.20944/preprints202303.0438.v1

Kuzman, T., Ljubešić, N. y Mozetič, I. (2023). ChatGPT: Beginning of an End of Manual Annotation? Use Case of Automatic Genre Identification. *arXiv preprint arXiv:2303.03953*. https://doi.org/10.48550/arXiv.2303.03953

Laerd Statistics (2019). Fleiss' kappa using SPSS Statistics. *Statistical tutorials and software guides*. Recuperado el 23 de abril de 2023, de https://statistics.laerd.com/

Lee, E. (2023). Is ChatGPT a False Promise?. United States of America. Recuperado el 8 de abril de 2023, de https://policycommons.net/artifacts/3528039/is-chatgpt-a-false-promise/4328874/

López Rodríguez, C. I. (2020). Marcos predicativos asociados al concepto signo y síntoma en textos sobre medicina en español. *Revista signos, 53*(103), 392-418. https://dx.doi.org/10.4067/S0718-09342020000200392

Lund, B. D. y Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*, *40*(3), 26-29. https://doi.org/10.1108/LHTN-01-2023-0009

Marimon, M. (2015). Tibidabo: a syntactically and semantically annotated corpus of Spanish. *Corpora 18*(1). https://doi.org/10.3366/cor.2015.0077

Marimon, M. (2015). Tibidabo: a syntactically and semantically annotated corpus of Spanish. *Corpora, 10*(3), 259-276. Recuperado el 15 de junio de 2022, de https://www.euppublishing.com/doi/abs/10.3366/cor.2015.0077

Martí Antonín, M. A., Taulé Delor, M., Màrquez i Villodre, L. y Bertran, M. (2007). Anotación semiautomática con papeles temáticos de los corpus CESS-ECE. *Procesamiento del Lenguaje Natural, (38)*, 67-76. Recuperado el 15 de junio de 2022, de https://www.redalyc.org/pdf/5157/515751738009.pdf

Moeller, S. y Hulden, M. (marzo de 2021). Integrating Automated Segmentation and Glossing into Documentary and Descriptive Linguistics. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, *1*, pp. 86-95). Recuperado el 19 de mayo de 2023, de https://journals.colorado.edu/index.php/computel/article/view/965/895

Newman, J. y Cox, C. (2020). Corpus Annotation. In M. Paquot y S. T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 25-48). Springer International Publishing.

Oh, B. D., y Schuler, W. (2023). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, *11*, 336-350. https://doi.org/10.1162/tacl_a_00548

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., y Alonso, A. (2023). Linguistic ambiguity analysis in ChatGPT. *arXiv e-prints*, arXiv-2302. https://doi.org/10.48550/arXiv.2302.06426

Periñán Pascual, J. C. (2012). En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica. *Onomázein: Revista de Lingüística, Filología y Traducción,* (26), 13-48. https://doi.org/10.7764/onomazein.26.01

Pichai, S., y Hassabis, D. (2023, December 6). Introducing Gemini: Our largest and most capable AI model. *Google The Keyword*. https://blog.google/technology/ai/google-gemini-ai/#sundar-note

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., . . . Wu, Q. M. J. (2023). A Review of Generalized Zero-Shot Learning Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4051-4070. https://doi.org/10.1109/TPAMI.2022.3191696

Pustejovsky, J. y Stubbs, A. (2013). *Natural Language Annotation for Machine Learning* (First ed.). O'Reilly.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M. y Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver?. *arXiv preprint arXiv:2302.06476*. https://doi.org/10.48550/arXiv.2302.06476

Reiss, M. V. (2023). Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. *arXiv preprint arXiv:2304.11085*. https://doi.org/10.48550/arXiv.2304.11085
Reyes, José A., Montes, Azucena, González, Juan G. y Pinto, D. E. (2013). Clasificación de roles semánticos usando características sintácticas, semánticas y contextuales. *Computación y Sistemas, 17*(2), 263-272. Recuperado el 02 de septiembre de 2021, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462013000200016&lng=es&tlng=es

Spinde, T. (diciembre de 2021). An interdisciplinary approach for the automated detection and visualization of media bias in news articles. In *2021 International Conference on Data Mining Workshops (ICDMW)*(pp. 1096-1103). IEEE. Recuperado el 19 de mayo de 2023, de https://ieeexplore.ieee.org/document/9679994

Sun, X., Dong, L., Li, X., Wan, Z., Wang, S., Zhang, T., ... y Wang, G. (2023). Pushing the limits of CHATGPT on NLP tasks. *arXiv preprint arXiv:2306.09719*. https://doi.org/10.48550/arXiv.2306.09719

Thapa, S., Naseem, U. y Nasim, M. (2023). From humans to machines: can ChatGPT-like LLMs effectively replace human annotators in NLP tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media* Association for the Advancement of Artificial Intelligence (AAAI). https://doi.org/10.36190/2023.15

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*. https://doi.org/10.48550/arXiv.2304.06588

Üstün, A., Bisazza, A., Bouma, G. y van Noord, G. (2020). UDapter: Language Adaptation for Truly Universal Dependency Parsing. *arXiv preprint arXiv:*2004.14327. https://doi.org/10.48550/arXiv.2004.14327

Vàzquez, G., Alonso, L., Capilla, J. A., Castellón, I. y Fernández, A. (2006). SenSem: sentidos verbales, semántica oracional y anotación de corpus. *Procesamiento del Lenguaje Natural* (37), 113-119. Recuperado el 15 de junio de 2022, de https://www.redalyc.org/pdf/5157/515751737015.pdf

Wallis, S. y Nelson, G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery, 5*(4), 305-335. https://doi.org/10.1023/A:1011453128373

Wang, Z., Xie, Q., Ding, Z., Feng, Y., y Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*. https://doi.org/10.48550/arXiv.2304.04339

Wu, Y. (2024). Evaluating ChatGPT: Strengths and Limitations in NLP Problem Solving. *Highlights in Science, Engineering and Technology*, *94*, 319-325. https://doi.org/10.54097/z15ne349

Zhang, B., Ding, D. y Jing, L. (2022). How would Stance Detection Techniques Evolve after the Launch of ChatGPT?. *arXiv preprint arXiv:2212.14548*. https://doi.org/10.48550/arXiv.2212.14548

Zhong, Q., Ding, L., Liu, J., Du, B. y Tao, D. (2023). Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. *arXiv preprint arXiv:2302.10198*. https://doi.org/10.48550/arXiv.2302.10198

Zhu, Y., Zhang, P., Haq, E. U., Hui, P., *y* Tyson, G. (2023). Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*. https://doi.org/10.48550/arXiv.2304.10145

Asking it directly: Large language models and corpus linguistics responses to language use questions

João Pedro Padua

Abstract

Until recently, the only way to empirically assess patterns of naturally-occurring language use was through query methods collectively referred to as corpus linguistics ("CL"). CL is an observational field that evolved through the refinement of methods for measuring different aspects of the frequency of linguistic units in the data (Gries, 2010). Although it provided a systematic way to test hypotheses about language use, one limitation of CL was that it had to limit itself mostly to "exposure rates" (Wallis, 2021, p. 47), which are, by definition, proxy variables: the corpus answers the question "given the text captured for assembling the corpus, is it common to find *x*?", not "given *x*, could it have been commonly generated by the subjects whose text went into the corpus?".

Large language models ("LLMs") are also trained on large corpora, but they transform these training data into billions of numerical parameters, stored in high-dimensional matrices that encode the data as knowledge. After training, LLMs can be asked questions directly, and their answers will appear as next-token predictions, conditioned on the knowledge it had stored and the question asked of them.

The research question in this paper is whether LLMs can be used as a complement to, or even *in lieu of*, corpus-assisted analysis, when the issue is how people will evaluate patterns of language use in specific settings.

To test this, I have used three of the state-of-the-art LLMs to replicate the analysis done by Gries and Slocum (2017) of the interpretation of a US legal statute that is at issue in the case *Smith v. U.S.*, decided by the US Supreme Court in 1993. The Court had to decide whether offering a machine gun as a barter for cocaine fell within the ordinary meaning of the verb phrase USES A GUN, which, according to the statute, constituted the basis for a sentence enhancement. A majority of the Court decided that it did. Using clever and complex queries to a reference corpus, Gries and Slocum seemed to have shown that the Court was empirically wrong. However, when I asked a synthetic sample of LLMs the same question, all but one of them agreed with the Court's majority.

The potential reasons and implications of this finding for linguistic research and CL in particular are discussed in the paper.

Keywords: corpus linguistics, artificial intelligence, large language models, statutory interpretation, synthetic samples

References

Gries, S. (2010). Useful statistics for corpus linguistics. In Sanchez, A. and Almela, M., editors, *A mosaic of corpus linguistics: selected approaches*, chapter 4, pages 269–291. Peter Lang, Frankfurt am Main

Gries, S. T. and Slocum, B. (2017). Ordinary Meaning and Corpus Linguistics. *BYU Law Review*, 2017(6):1417–1471.

Wallis, S. (2021). *Statistics in corpus linguistics research: A new approach*. Routledge, New York.

Stefan Gries

Towards prototypes/ordinary meaning in legal interpretation: High-cue validity features from AI and synchronic corpus data

One of the most important notions in legal interpretation is that of ordinary meaning, the notion that courts need to interpret terms that are used, but not defined, in constitutions, laws, contracts, and other legal contexts according to "what those words would mean in the mouth of a normal speaker of English, using them in the circumstances in which they were used" (Holmes 1899:417). To a linguist, this formulation (and others like it) probably triggers an association to the notion of prototype meaning and indeed many legal scholars have approached ordinary meaning with corpus methods reminiscent of how a cognitive or usage-based linguist might approach the identification of prototypes based on corpus data.

While such work is generally promising, it has often been simplistic, at least from the point of usage-based and corpus linguists: prototypicality and category membership – e.g., is an airplane a vehicle? – was basically 'operationalized' on the basis of co-occurrence frequency in concordance lines, which is problematic for at least two reasons: First, such work is based on extensionalist semantics of co-occurrence rather than how prototype theory defines prototypicality/category membership (on the basis of defining features with high cue validities, i.e. features that are exhibited my most category members but hardly anything else. Second, such work neglects the different reasons underlying co-occurrence: tires is not a collocate of vehicle even though the prototypical vehicles has tires.

This paper exemplifies a combination of AI use and corpus data to approach prototypicality and category membership. It uses Perplexity and DeepSeek to identify high-cue validity defining features for categories and uses them to quantify degrees of category membership based on distributional-semantics models trained on corpora (like COHA) in a way that is nearly fully bottom-up and, important in legal contexts, nearly impossible to game in favor of a preferred legal outcome/interpretation.

Theoretical and Practical Approaches to the Interface between AI and Corpus Linguistics
Tobias Bernaisch

Corpus linguistics displays deep-rooted connections with generative AI as a currently prominent form of artificial intelligence. Just as linguistic corpora are large collections of authentic texts representative of a particular type of language use (cf. Mukherjee 2002), text-based AI models rely on large collections of texts as well. Still, the expectably increasing adoption of generative AI tools in business, educational and creative settings affects the work of corpus linguists both on theoretical and practical levels. The central research questions of this paper are, thus, a) whether – on a theoretical level – the integration of AI texts into linguistic corpora is reconcilable with existing definitions of corpora and b) how helpful – on a practical level – current AI tools can be in corpus-linguistic workflows.

Regarding corpus-linguistic theory, the implications of the emergence of generative AI are discussed with regard to authenticity and representativeness. Whether corpus linguists consider AI texts authentic, i. e. "language produced for the purpose of communication, not for linguistic analysis" (Stefanowitsch 2020: 23), will influence the inclusion of AI texts in a linguistic corpus. This decision will have implications for the representativeness, i. e. the degree to which a linguistic corpus as a sample of texts mirrors the entirety of texts in a population, of corpora in the AI era. In this light, arguments for and against the integration AI-generated texts in linguistic corpora are offered.

Practically, AI tools can support corpus-linguistic work, but how accurate and consistent are the respective results? With a view to this question, a corpus-linguistic work routine for the dative alternation (cf. e. g. Röthlisberger 2018) including data extraction, export, cleaning and annotation has been implemented with ChatGPT and critically evaluated. The results of this AI-aided corpus-linguistic work routine are promising, but also leave room for improvement.

References

Mukherjee, J. 2002. *Korpuslinguistik und Englischunterricht: Eine Einführung*. Peter Lang.

Röthlisberger, M. 2018. *Regional Variation in Probabilistic Grammars: A Multifactorial Study of the English Dative Alternation*. PhD thesis: KU Leuven.

Stefanowitsch, A. 2020. *Corpus Linguistics: A Guide to the Methodology*. Language Science Press.

Human versus AI? – Harmonising human and artificial intelligence for corpus linguistics

Andreas Weilinghoff

Ever since the successful implementation of transformer models with attention mechanisms (Vaswani et al. 2017), numerous fields in Natural Language Processing (NLP) have seen significant improvements. While much attention has been given to machine translation and chatbot assistants (e.g. ChatGPT, Gemini, DeepSeek), speech recognition has also drastically improved in recent years (Jurafsky and Martin 2025).

This plenary talk explores the processing of spoken data in corpus linguistics. It provides an overview of how phonetic research and speech technology have evolved over the centuries – from Edison's phonograph to Amazon Alexa and beyond. Particular emphasis will be placed on the latest cutting-edge developments brought about by recent AI-based tools. These tools offer new possibilities for corpus linguistics, especially in the automatic processing and transcription of spoken data.

A central part of the talk is a study comparing the speed and accuracy of human transcribers with the latest end-to-end automatic speech recognition (ASR) models. I will discuss to what extent AI can support or even replace human efforts in corpus preparation tasks and where its current limitations lie. Drawing on a reference study of various English varieties, including the ICE Nigeria (Wunder et al. 2008) and ICE Scotland (Schützler et al. 2017) corpora, I will demonstrate how a hybrid approach – combining human expertise and artificial intelligence – can yield the most efficient and accurate results for transcription and corpus compilation.

References:

Jurafsky, D. & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. https://web.stanford.edu/~jurafsky/slp3/

Schützler O., Gut U. & Fuchs R. (2017). New perspectives on Scottish Standard English. Introducing the Scottish component of the International Corpus of English. In Hancil, S., Beal, J. (Eds.), *Perspectives on Northern Englishes* (pp. 273–301). Berlin: De Gruyter Mouton.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). *Attention Is All You Need*. https://arxiv.org/pdf/1706.03762.pdf

Wunder E.-M., Voormann H. & Gut, U. (2010). The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal*, 34, 78–88.

Large Language Models and the Process of Corpus Linguistic Research

Martin Klotz, Thomas Krause, Anke Lüdeling, Carolin Odebrecht, Malte Dreyer
(all Humboldt-Universität zu Berlin)

Can Large Language Models (LLMs) efficiently support existing corpus building and analysis workflows and at what epistemological cost? What do we gain and lose by relying (solely) on LLMs in our workflows?

LLMs have proven capable of solving many tasks by mere instruction rather than additional fine-tuning (Brown et al., 2020). This questions the necessity of specialized annotation tools and manual post-correction for doing corpus-linguistic research. Even directly obtaining linguistic characterizations from LLMs seems thinkable.

In our work we compare manual core linguistic annotations with such generated by an LLM (Jiang et al., 2024). In the same manner, we explore the application of a novel register-related annotation scheme (Lehmann, to appear, pp. 217–226) to test the LLM's capabilities, as no annotations or documentation were part of its training data. We conduct our study using the NoSta-D corpus (Dipper et al., 2013), which provides manual annotations for spoken language, learner language, historical data, literary prose, and a reference set of newspaper data.

Manual annotation is a vital part of the research process and category building is considered its main deliverable as a reflection of a deeper scientific understanding (Shadrova et al., 2025). In our analyses, we therefore focus on challenging phenomena covered by NoSta-D's syntactic dependency scheme and the register analysis. This way, we can examine whether LLM-based annotation processes can yield equal or new fine-grained distinctions or insights, or at least show promising performance.

We complement such layer-wise annotation with an LLM approach to obtain linguistic characterizations directly from primary data, i. e. from scans of the historical data (using Bai et al., 2025) and from recordings in the dialogue subcorpus. In our contrastive analysis, we highlight the differences in detailed explanations, transparency, and understanding of the linguistic problems and question, whether prompting vs. annotation generates comparable linguistic insights and epistemological value.

Finally, we incorporate in our assessment the cost of prompt engineering and reproducibility as another core aspect of research quality.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901, Vol. 33). Curran Associates, Inc. https:// proceedings. neurips. cc/ paper_ files/ paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Jiang, A., Sablayrolles, A., Tacnet, A., Kothari, A., Roux, A., Mensch, A., Herblin-Stoop, A., Garreau, A., Birky, A., Bam4d, Bout, B., de Monicault, B., Savary, B., Rambaud, C., Feldman, C., Chaplot, D. S., de las Casas, D., Costa, D., Arcelin, E., Hanna, E. B., Metzger, E., Blanchet, G., Lengyel, G., Bour, G., Lample, G., Rajaona, H., Roussez, H., Sattouf, H., Mack, I., Delignon, J.-M., Chudnovsky, J., Murke, J., Khandelwal, K., Stewart, L., Martin, L., Ternon, L., Saulnier, L., Lavaud, L. R., Jennings, M., Pellat, M., Torelli, M., Lachaux, M.-A., Janiewicz, M., Seznec, M., Schuhl, N., Muhs, N., de Garrigues, O., von Platen, P., Jacob, P., Buche, P., Reddy, P. K., Savas, P., Stock, P., Sauvestre, R., Vaze, S., Subramanian, S., Garg, S., Yang, S., Antoniak, S., Scao, T. L., Schueller, T., Lavril, T., Wang, T., Gervet, T., Lacroix, T., Nemychnikova, V., Shang, W., Sayed, W. E., & Marshall, W. (2024). Mistral-large-instruct-2411. https://huggingface.co/mistralai/Mistral-Large-Instruct-2411

Lehmann, N. (to appear). *The intricacies of register variation across languages* [Doctoral dissertation, Humboldt-Universität zu Berlin]. Retrieved July 30, 2025, from https://box.hu-berlin.de/f/775e5da654ef4b9fa2b6/?dl=1

Dipper, S., Lüdeling, A., & Reznicek, M. (2013). NoSta-D: A Corpus of German Non-Standard Varieties. In M. Zampieri & S. Diwersy (Eds.), *Non-Standard Data Sources in Corpus-Based Research* (pp. 69–76). Shaker Verlag.

Shadrova, A., Lüdeling, A., Klotz, M., Hartz, R. G., & Krause, T. (2025). „Step away from the Computer!" – Über die linguistische Datenkategorisierung als Erkenntnisprozess und daraus folgende Herausforderungen bei der Nachnutzung von Annotationen und Annotationstools. *Zeitschrift für germanistische Linguistik*, *53*(1), 166–214. https://doi.org/doi:10.1515/zgl-2025-2005

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., & Lin, J. (2025). Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Towards a deep-learning-powered POS-tagger of beginner and intermediate EFL interlanguage

Valentin Werner & Sabine Weber (University of Bamberg)

Part-of-speech (POS) tagging of spoken and written language is a fundamental annotation/sequencing task in natural language processing (NLP), enhancing precision and recall in specific subsequent corpus queries and facilitating automatic annotation. Traditional POS taggers often rely on rule-based or statistical methods, which struggle with the non-standard patterns and disfluencies found in learner interlanguage (Van Rooy & Schäfer 2002). While recent advancements in neural networks and deep learning have improved POS tagging accuracy (Chiche & Yitagesu 2022), many models are trained on native speaker corpora (and often on written material from more formal registers), neglecting idiosyncratic spellings, omissions, and transfer-induced errors common in learner language (Meurers & Dickinson 2017; Tammekänd & Torn-Leesik 2023). In addition, while there have been some recent attempts to adapt POS taggers to the specificities of interlanguage (e.g. Nagata et al. 2018), they are commonly trained on interlanguage by advanced learners at higher proficiency levels (typically university students). Thus, a significant gap in systems for beginner and intermediate learners persists.

To address this lacuna, this interdisciplinary presentation reports on a linguistically informed AI-based framework for POS tagging of beginner and intermediate EFL learner data. Key research questions include: (1) How do state-of-the-art rule-based POS taggers handle EFL interlanguage, and what are their limitations? (2) How do AI-driven POS taggers compare against manually annotated gold standards and rule-based systems? (3) How can NLP models be enhanced to capture specific error patterns, morphosyntactic variability, or the presence of multiple languages in EFL learner language?

Using a mixed-methods approach, this project combines manual and automated annotation techniques. A representative subset of the YGLE corpus that contains data from beginner and intermediate EFL learners (YGLE 2025) will be manually annotated to create a reliable gold standard dataset. This dataset will serve to train advanced deep learning algorithms, particularly transformers like BERT, for context-aware tagging. Model performance will be evaluated based on precision, recall, and F1-score metrics against the gold standard, revealing insights into the differences between automated and expert annotations.

References

Chiche, A., & Yitagesu, B. (2022) Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*, 10. https://doi.org/10.1186/s40537-022-00561-y

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, *67*(s1), 66–95. https://doi.org/10.1111/lang.12233

Nagata, R., Mizumoto, T., Kikuchi, Y., Kawasaki, Y., & Funakoshi, K. (2018). A POS tagging model designed for learner English. In W. Xu, A. Ritter, T. Baldwin, & A. Rahimi (EDS.), *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (pp. 39–48). Brussels: ACL.

Tammekänd, L., & Torn-Leesik, L. (2023). Automatic part-of-speech tagging of the Tartu Corpus of Estonian Learner English with CLAWS7. *Taikomoji kalbotyra*, *20*, 121–135. https://doi.org/10.15388/Taikalbot.2023.20.9

Van Rooy, B., & Schäfer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, *20*(4), 325–335. https://doi.org/10.2989/16073610209486319

YGLE. (2025). *The Corpus of Young German Learner English*. https://www.ygle.de

Perspectivity in Legal Texts. A comparative analysis of traditional corpus linguistic software and LLMs

Abstract for Conference “Corpus linguiscs in the AI Era”, JLU Gießen, May, 07 – 09 2026

Margret Mundorf, Heidelberg

In my proposal I will present a section of the ongoing PhD project in the field of legal linguistics at Heidelberg University. It investigates how extra-linguistic facts are represented linguistically in different types of legal texts. The description of facts is an essential preliminary to the legal assessment. These descriptions are employed to transform extra-linguistic reality into legal cases and decisions, which in turn exert an influence on reality. These linguistically produced descriptions of facts (Felder/Vogel 2017) are subject to fundamental perspectivization processes (Köller 2004). Perspectivization patterns and perspectivization strategies are used as a heuristic concept to describe linguistic mediation strategies in law (Mundorf 2024). In specialized domains that demand domain-specific expertise, such as law, the employment of specialized models may be imperative. Small Language Models (SLMs) are particularly well-suited for their low inference latency, effectiveness in terms of costs and development, and ease of customization and adaptability (Wang et al. 2024).

AI is gaining traction in legal practice, within law firms and the judiciary. Research conducted by Wegerhoff (2025) demonstrates that LLMs exhibit a reduced capacity to articulate linguistic vagueness. Nevertheless, they do permit an evaluation of the quality of factual descriptions, such as those provided by experts in family law proceedings. A small corpus of legal texts will be selected to compare traditional corpus linguistic software (AntConc) with various proprietary and open LLMs with regard to aspects of vagueness, evidentiality, and modality. The proposed method integrates corpus linguistic approaches with a range of proprietary and open-source models. These components are to be trained in a decentralized manner specifically for legal applications, leveraging data-driven analyses augmented with expert knowledge from court decisions and decision-making data from court files in family law. The objective is to ensure the method’s usability for a test phase.

References

Brodowski, Dominik (2024): Datengestützte Prognose justizieller Entscheidungen. In: Liane Wörner, Rüdiger Wilhelmi, Jochen Glöckner, Marten Breuer und Svenja Behrendt (Hrsg.): Digitalisierung des Rechts: de Gruyter, S. 125–142.

Bubenhofer, Noah (2024): Die Lektüre von Texten und Daten. Data Philology statt Data Science. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 54 (2), S. 269–283. DOI: 10.1007/s41244-024-00338-1.

Dahl, Matthew; Magesh, Varun; Suzgun, Mirac; Ho, Daniel E. (2024): Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. In: *Journal of Legal Analysis* 16 (1), S. 64–93
Felder, Ekkehard (2008): Das Forschungsnetzwerk „Sprache und Wissen“: Zielsetzung und Inhalte. In: Zeitschrift für Germanistische Linguistik 36 (2), S. 270–276.

Felder, Ekkehard; Vogel, Friedemann (2017): Einleitung. In: Ekkehard Felder und Friedemann Vogel (Hg.): Handbuch Sprache im Recht. Berlin, Boston: de Gruyter (Handbücher Sprachwissen, 12), S. IX–XIX.

Hoffmann-Riem, Wolfgang (2022): Der Umgang mit Wissen bei der digitalisierten Rechtsanwendung. In: Ino Augsberg und Gunnar Folke Schuppert (Hg.): Wissen und Recht. Baden-Baden: Nomos (Interdisziplinäre Studien zur Wissensgesellschaft, Band 1), S. 509–560.

Holste, Alexander (2024): Automatisierte Wissenskommunikation. Berlin: Frank & Timme.

Ji, Ziwei; Lee, Nayeon; Frieske, Rita; Yu, Tiezheng; Su, Dan; Xu, Yan et al. (2023): Survey of Hallucination in Natural Language Generation. In: ACM Computing Surveys 55 (12), S. 1–38.

Köller, Wilhelm (2004): Perspektivität und Sprache. Zur Struktur von Objektivierungsformen in Bildern, im Denken und in der Sprache. Berlin: de Gruyter.

Mundorf, Margret (2024): Recht vermitteln. Perspektivität in der Vermittlung juristischen Wissens in Fort- und Weiterbildung. In: Katja Leyhausen-Seibert, Anna Menzel und Friedemann Vogel (Hrsg.): Wissen in Recht und Sprache – Viele Stimmen, vage Grenzen. Berlin: Duncker & Humblot (Sprache und Medialität des Rechts), S. 257–285.

Mundorf, Margret (2024): Legal Linguistic Memos mit Large Language Models. Automatisierte Erfassung und Klassifizierung von Sachverhaltsbeschreibungen im Familienrecht. (Wissenschaftliches Poster: Best Poster Award der 3. Text+ Plenary zum Thema "Große Sprachmodelle (LLMs) und deren Nutzung). Online verfügbar unter https://events.gwdg.de/event/638/contributions/2744, zuletzt geprüft am 10.10.2024.

Polanyi, Michael (1985): Implizites Wissen. Frankfurt am Main: Suhrkamp (Suhrkamp-Taschenbuch Wissenschaft, 543).

Scharloth, Joachim; Bubenhofer, Noah (2012): Datengeleitete Korpuspragmatik. Korpusvergleich als Methode der Stilanalyse. In: Ekkehard Felder, Marcus Müller und Friedemann Vogel (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin: de Gruyter (Linguistik - Impulse & Tendenzen, 44), S. 194–230.

Speer, Robyn (2024): Note from September 2024: Why wordfreq will not be updated. Online verfügbar unter https://github.com/rspeer/wordfreq/blob/master/SUNSET.md?ref=404media.co, zuletzt aktualisiert am 23.09.2024.

Vogel, Friedemann (2024): Der Richter, (k)ein Bot?! In: Liane Wörner, Rüdiger Wilhelmi, Jochen Glöckner, Marten Breuer und Svenja Behrendt (Hrsg.): Digitalisierung des Rechts: De Gruyter, S. 9–26.

Wang, Fali; Zhang, Zhiwei; Zhang, Xianren; Wu, Zongyu; Mo, Tzuhao; Lu, Qiuhao et al. (2024): A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. Online verfügbar unter http://arxiv.org/pdf/2411.03350v2.

Wegerhoff, Dennis (2025): Uncanny Semantics: How AI and Human Authors Use Language Differently in Academic Writing. This preprint is made available for public access. A revised version will be prepared for future submission to a peer-reviewed journal. Online verfügbar unter https://zenodo.org/records/15600548.

Wu, Kevin; Wu, Eric; Zou, James (2024): How faithful are RAG models? Quantifying the tug-of-war between RAG and LLMs' internal prior. Online verfügbar unter https://doi.org/10.48550/arXiv.2404.10198.

Title: Exploring the Capability of Gemma4-E4B for Annotation of Proximization Strategies

Xiaoyun Huang, Mo Han

Abstract: This study explores the capability of small-scale large language models for automated annotation in political discourse analysis. Using Gemma (Gemma4-E4B) as the base model and Proximization Theory as the analytical framework, the study aims to annotate German parliamentary debates on the Russia–Ukraine war in order to investigate how German parliamentarians discursively construct Russia as a threat. The research was conducted in three phases. First, 100 parliamentary texts were manually annotated as a gold standard corpus. Second, three prompt versions were developed to guide the model in annotating an additional 100 texts, allowing for a comparison of different prompt engineering strategies. Third, 80 texts were used for fine-tuning, while 20 texts were reserved for validation. The findings show that iterative prompt optimization resulted in only marginal performance improvements. They suggest that prompt engineering alone faces inherent limitations when applied to small-scale models in complex discourse annotation tasks. Fine-tuning, by contrast, substantially improved precision, producing more accurate and less noisy annotations, although it reduced recall by making the model more conservative and less willing to annotate uncertain instances. In general, the precision gains achieved through fine-tuning exceeded those obtained through prompt iteration, indicating that task-specific adaptation provides greater practical value than prompt optimization alone for proximization annotation. Future research will focus on expanding the fine-tuning dataset, exploring larger models such as Gemma 27B and Llama 3-8B, and incorporating active learning strategies to better balance precision and recall.

Keywords: Proximization Theory; Political Discourse Analysis; Large Language Model; Gemma; Text Annotation; Fine-tuning

Automated Target Hypothesis Generation in German Learner Corpora Using LLMs

Torsten Zesch (Fernuniversität Hagen) & Katrin Wisniewski (Universität Leipzig)

Understanding a second or foreign (L2) learner production involves a combination of interpretation and (re)generation and therefore is subject to ambiguity and variation. It is thus good research practice to explicate the understanding in form of a target hypothesis (Lüdeling, 2008; Reznicek et al., 2013), i.e., a normalized version of what the learner meant to say. As the term "hypothesis" suggests, there is not necessarily a single L1-like reference version that presents itself as "correct", but target hypotheses are specific to the analysis to be performed.

While target hypotheses are very important for conducting corpus-based analyses of language acquisition patterns, only a tiny fraction of learner corpora come with manually annotated target hypotheses due to the high costs of creating them. LLMs have the potential to revolutionize learner corpus research by largely automating the annotation process.

In our talk, we will critically reflect on the capabilities of LLMs in generating target hypotheses. We will also present evidence from using automatically generated target hypotheses within the scope of the DAKODA project (Federal Ministry of Research, Technology and Space, 10/2022-09/2022), where we developed methods for automatically annotating developmental stages of verb placement as articulated by Processability Theory (e.g., Pienemann, 1998; Lenzing et al., 2019). We will explain which role LLM-generated target hypotheses played in the process.

Lenzing, A., Nicholas, H., & Roos, J. (2019). *Widening contexts for Processability Theory: Theories and issues*. Benjamins.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In M. Walter & P. Grommes (Hrsg.), *Fortgeschrittene Lernervarietäten* (S. 119–140). Max Niemeyer Verlag.

Pienemann, M. (1998). *Language processing and second language development. Processability theory*. Benjamins.

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C., & Andreas, T. (2012). *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Institut für deutsche Sprache und Linguistik, HU Berlin. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/standardseite

Large language models in Metadiscourse Identification

Man Zhang

Large language models (LLMs) exhibit remarkable capabilities in understanding and generating natural human language. We believe that they, with exceptional linguistic abilities, could assist in linguistic research. The present study focuses on the LLM-powered metadiscourse research.

Metadiscourse is a contextual phenomenon (Mauranen 1993; Hyland 2005; Ädel 2006; Zhang 2022). The identification of metadiscourse markers in texts need to be done manually, which is time-consuming and labor-intensive. This limits the scope of metadiscourse research. In order to improve the efficiency of metadiscourse research, and thus investigate metadiscourse more comprehensively, we intend to automate metadiscourse identification in texts with the help of LLMs.

Taking research article abstracts as an example, we built a corpus of English abstracts and used ChatGPT and BERT to identify metadiscourse markers in the corpus automatically. Main research questions are: (1) How can LLMs be used in metadiscourse identification in abstracts? (2) How good are LLMs for the automatic metadiscourse identification in abstracts? For the work with ChatGPT, we mainly provided it with the following prompts: the definition and identification criteria of metadiscourse and some examples of metadiscourse markers. With the identification results, we adjusted the prompts. This process was repeated for several rounds until ChatGPT could do a satisfactory job (F1 score more than 0.80). For the BERT-powered automatic identification of metadiscourse, we fine-tuned the model for the specific purpose of metadiscourse identification on a manually-annotated corpus which is a subsection of the corpus. After repeated-rounds of model fine-tuning, we got a model with good performance (F1 score more than 0.80).

Both ChatGPT and fine-tuned BERT could automatically identify metadiscourse markers in abstracts satisfactorily. ChatGPT is limited in handling large corpora and analyzing complicated contexts. The fine-tuned BERT model could overcome the limitations to a certain degree. Researchers, however, need some programming knowledge and hyperparameter tuning techniques.

References

Ädel, A. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hyland, K. 2005. *Metadiscourse: Exploring Interaction in Writing*. London/New York: Continuum.

Mauranen, A. 1993a. 'Contrastive ESP rhetoric: Metatext in Finnish–English economics texts,' *English for Specific Purposes* 12 (1): 3–22.

Mauranen, A. 1993b. *Cultural Difference in Academic Rhetoric.* Frankfurt: Peter Lang.

Zhang, M. 2022. 'Variation in metadiscourse across speech and writing: A multidimensional study,' *Applied Linguistics* 43 (5): 912–33.

Linguistic subtlety at scale: AI-assisted annotation of Mandarin grammar

Author: Ruiming Ma[1]

Abstract

One of the central challenges in linguistic annotation involves integrating contextual clues to disambiguate grammatical variants (Schneider et al. 2018). This is particularly salient in Mandarin Chinese, where syntactic alternations often hinge on discourse constraints (LaPolla 1990). This study investigates the potential of large language models (LLMs) to support the annotation of such variation, focusing, among other, on the interchangeability between the *bǎ*-disposal construction and the canonical SVO word order (Fang & Liu 2021), see (1).

(1)	把	佐料	一起	和	在	鱼	里面
	bǎ	***zuǒliào***	*yìqǐ*	*huó*	*zài*	*yú*	*lǐmiàn*
	BA	**seasoning.O**	together	mix.V	LOC	fish	inside

"Mix the seasonings with the fish."

ANNOTATION	Bǎ is not interchangeable with SVO word order.
REASONING	The location complement (LOC-fish-inside) must follow the verb (mix) directly. In SVO, the object (seasoning) would block this.

We focus on (1) the **ACCURACY** of LLMs (at this stage: OpenAI's ChatGPT-4o vs. DeepSeek- R1); (2) the **CONSISTENCY** of the generated annotations, especially with regard to shifting contextual clues. Evaluation involves the comparison between generated annotation output and human annotation through agreement rates and context-sensitive error patterns.

Data come from the large-scale HKUST Mandarin Telephone Speech Corpus (149 hours of conversational telephone speech). With the current models at moment of writing, we find that both models demonstrate promising results in automating labour-intensive annotation. However, DeepSeek aligns closer with human annotations, particularly in distinguishing subtle alternations. This performance gap highlights differences in model design and training that impact linguistic interpretation. Tentatively, this is related to the handling of Mandarin data. We are currently further investigating how to improve performance for replicability, transparency, and linguistic validity, aiming to benefit research communities involved in such annotation tasks.

References

Fang, Y., & Liu, H. (2021). Predicting syntactic choice in Mandarin Chinese: A corpus-based analysis of ba sentences and SVO sentences. Cognitive Linguistics, 32(2), 219-250.

LaPolla, R. J. (1990). Grammatical relations in Chinese: Synchronic and diachronic considerations. University of California, Berkeley.

---

[1] Department of Linguistics, KU Leuven, Belgium

Liu, Y., Fung, P., Yang, Y., CIERI, C., Huang, S., & Graff, D. (2006). HKUST/MTS : A very large scale mandarin telephone speech corpus. Lecture Notes in Computer Science, 724–735. Schneider N, Hwang JD, Srikumar V, et al. Comprehensive supersense disambiguation of English prepositions and possessives. In: Proceedings of the 56th Annual meeting of the Association for Computational Linguistics, 2018, vol. 1: Long Papers, pp.185–196.
Melbourne: Association for Computational Linguistics. v1/P18-1018. https://aclanthology.org/P18-1018

Sociolinguistic competence in an AI-generated corpus: A study of *that* complementizer

Emily Herman / Pablo Requena

Given AI agents' fluency as well as the fact that they are trained using real language samples, it is easy to assume human-like language ability. Yet, little is known about whether these agents replicate sociolinguistic variation present in human language [1,2,3].

To test the sociolinguistic patterns of variation in AI-generated language, we expand on Duncan (2024) by analyzing a novel corpus of sociolinguistic interviews generated using AI. In English, complement clauses may variably occur with *that* or ∅ [4], as in 'I know ***that***/∅ teachers need supplies.' Crossdialectal sociolinguistic analysis has found this variation to be conditioned by internal linguistic factors, including the subject in the matrix and complement clauses, and the lexical verb in the matrix clause [5,6,7,8]. Our research question is: Do AI-generated speakers condition variation in human-like ways with complementizer that/∅ production?

We asked ChatGPT-4 [9] to generate 32 Labovian-style interviews [10] for speakers in Santa Barbara, California. Speakers were counterbalanced for gender (male & female), socioeconomic status (working & middle class), and age group. All variable *that*/∅ contexts were manually extracted and coded for the matrix verb and subject. We excluded the collocations 'I think…' and 'You know…' categorically associated with ∅ [8]. The remaining 787 tokens were analyzed using a *glmer* model fitted with gender, SES, and matrix verb and subject as independent variables [11,12,13].

Overall, there was a 27% rate of overt *that* production and the matrix verbs 'know', 'make sure', 'remember', 'say', 'think', and 'other' infrequent matrix verbs favor ∅ (see Table 1). These results are partially consistent with patterns observed in naturalistic corpora [5,6,7,8]. Additionally, although such research did not find an effect of gender or socio-economic status [5], ChatGPT-4 appears to impose these external conditioning factors on the sociolinguistic variation; male participants favor overt *that* and working class participants favor ∅**.**

References: **[1]** Duncan, D. (2024). Journal of Computer-Assisted Linguistic Research 8:51-75. ISSN: 2530-9455 **[2]** Lotze, N. *Human-Machine Interaction as a Complex Socio-Linguistic Practice*. **[3]** Kelly-Holmes, H. (2024). Artificial intelligence and the future of our sociolinguistic work. Journal of Sociolinguistics, 28(5), 3–10. https://doi.org/10.1111/josl.12678 **[4]** Storms, G. (1966). That-clauses in modern English **[5]** Gadanidis, T., Kiss, A., Konnelly, L., Pabst, K., Schlegl, L., Umbal, P., & Tagliamonte, S. A. (2023). Integrating qualitative and quantitative analyses of stance: A case study of English that/ zero variation. Language in Society, 52(1), 27–50. https://doi.org/10.1017/S0047404521000671 [6] Tagliamonte, S., & Smith, J. (2005). No momentary fancy! The zero 'complementizer' in English dialects. English Language and Linguistics, 9(2), 289–309. https://doi.org/10.1017/S1360674305001644 **[7]** Thompson, S. A., & Mulac, A. (1990). *He discourse conditions for the use of the complementizer that in conversational*. **[8]** Torres Cacoullos, R., & Walker, J. A. (2009). On the persistence of grammar in discourse formulas: A variationist study of that. *Linguistics*, *47*(1), 1–43. https://doi.org/10.1515/LING.2009.001 **[9]** OpenAI (2025). https://chatgpt.com/ **[10]** Labov, William (1984). Field methods of the project on linguistic change. In Language in Use: Readings in Sociolinguistics, John Baugh and Joel Sherzer (eds.), 28–54. Englewood Cli¤s, NJ: Prentice-Hall. **[11]** R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ **[12]** RStudio Team. (2020). RStudio: Integrated development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/ **[13]** Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4."

Table 1. Output of the logistics regression model. Reference level: that, gender = female, SES = middle class, matrix verb = believe.

	Estimate	*Std. error*	*p-value*
Intercept	0.54232	0.65972	0.411048
Gender			
male	0.93240	0.32306	0.0039 **
SES			
working class	-0.8653	0.3049	0.004543 **
Matrix subject			
you	0.35316	0.67976	0.603384
other	0.80225	0.2551	0.001661 **
Matrix verb			
think	-5.43649	1.1913	5.03e-06 ***
know	-2.86701	0.76268	0.000171 ***
tell	-1.43072	0.82061	0.081249 .
say	-3.53080	0.74931	2.45e-06 ***
remember	-4.87709	1.19613	4.55e-05 ***
be	-1.26365	0.67222	0.060133 .
make sure	-4.30659	0.98378	1.20e-05 ***
have	0.06162	0.74186	0.933802
other	-1.77911	0.64882	0.006106 **

Title of Presentation: Generative AI-Assisted Text Sentiment Analysis Applications

Research Questions:

1. **How can generative AI (specifically ERNIE 3.0 Base) be leveraged to create *semi-authentic corpora* for enhancing sentiment analysis systems that measure *emotional development* in K-12 education**
2. **To what extent can *AI-generated semi-authentic corpora* (controlled via keyword-guided synthesis) compensate for the scarcity of authentic educational texts and improve model predictive accuracy and expert consensus?**
3. **Can semi-authentic corpora serve as a *valid component* of linguistic corpora for pedagogical emotion research, and under what conditions?**

The attached abstract (under 300 words, excluding references) briefly introduces the study's rationale and key logic points; comprehensive research details will be elaborated in subsequent communications.

I confirm that this abstract is original and has not been published elsewhere. I look forward to the opportunity to present my findings at this timely forum.

Generative AI-Assisted Text Sentiment Analysis Applications

Generative AI offers significant speed and efficiency in building low-resource corpora. As a valuable linguistic research tool, it produces controllable, semi-authentic texts – an essential component of modern corpora.

Text sentiment analysis holds substantial potential in education, particularly for assessing emotional development objectives within teaching goals. This study develops an ERNIE 3.0 Base-based system to measure emotional development in K-12 students. However, challenges arise from insufficient digital resources in schools, scarcity of theme-specific texts, and the high cost of targeted data collection, resulting in inadequate training data. Generative AI effectively addresses this gap. By inputting sample texts, the AI generates large volumes of similar content. Controlled keyword input guides the creation of theme-specific texts for data augmentation. Larger models produce outputs with greater coherence, topic relevance, and standardization than some authentic texts, offering reasonable data reliability.

Critically, models trained on this semi-authentic data demonstrate significantly improved predictive performance and expert consistency compared to models trained on limited authentic data alone.

Zhao Wenjie

A Domain-Specific LLM for Error Annotation in German Learner Writing: Construction and Task-Based Evaluation

Yawen Zheng, Yuan Li

Abstract: Analyzing language errors in learner output provides insights into second language acquisition and is a primary motivation for building learner corpora. Error annotation encompasses identifying, categorizing, and correcting errors. However, current practices rely heavily on manual effort, making them time-consuming, labor-intensive, and prone to inconsistency. The emergence of large language models (LLMs) offers new possibilities for automating large-scale error annotation in learner writing.

Based on the *Chinesisches Deutschlernerkorpus* (CDLK) and following a systematic error annotation scheme, this study first manually annotated 3,400 learner texts. Using this annotated dataset, a generic LLM was trained via supervised fine-tuning (SFT) to develop *Dr. Write,* a domain-specific LLM for German writing feedback, with error annotation as a core functionality.

To evaluate the performance of *Dr. Write*, this study focuses on the following research questions: (1) How well does *Dr. Write* perform in error annotation? Are there performance variations across error types and texts with different error densities? (2) Compared to a generic LLM, how does the domain-specific LLM *Dr. Write* perform on error annotation tasks, and in what respects do their performances differ?

To answer these questions, we first selected 200 learner texts covering diverse error types and densities as test data. Next, the generic LLM ChatGPT-4o was chosen as a comparison model. To ensure task alignment, ChatGPT-4o was guided using prompts tailored to the same error annotation scheme. Finally, using manually annotated data as the benchmark, we compared the performance of *Dr. Write* and ChatGPT-4o across three dimensions: error identification, categorization, and correction, with particular attention to how each model performs on texts with different characteristics.

The study aims to empirically validate the applicability of LLMs for error annotation and explore the potential of supervised fine-tuning in improving the accuracy of error annotation.

Keywords: error annotation; domain-specific LLM**;** generic LLM; supervised fine-tuning (SFT)

How small can we go? Assessing the reliability of small open-weight LLMs for corpus annotation

Muhammad Shakir and Ellen Le Foll

Large Language Models (LLMs) are increasingly used in annotation and tagging tasks in corpus linguistics (e.g. Baker et al., 2025; Yu et al., 2024). Prior studies have generally tested very large LLMs like Open AI's GPT-4 or comparable models by Google. While these larger models may have superior capabilities, they raise serious environmental concerns, are expensive to run and thus inaccessible to many researchers, and, in many cases, cannot be used for personal data protection reasons and/or data leakage risks. Moreover, these large cloud-hosted LLMs typically do not allow researchers to control crucial hyper parameters like random seed, top_p, top_k, and temperature that can make LLM's outputs more predictable, and hence reproducible (Reiter, 2025).

In this paper we explore the use of smaller, locally-run open-weight models via ollama and an RTX 4090 for corpus linguistic annotation. We compare the outcomes of annotation tasks of varying grammatical and/or semantic complexity. In our first experiment, we checked for explicit information about country of origin in newspaper comments to classify each post as *local* or *foreign*. We compared the outputs of the following open-weight models: Qwen3.14b.q8_0, Qwen3.30b.a3b.q4_K_M, Deepseek.R1.14b.qwen.distill.q8_0, Gemma3.12b.it.q8_0, Deepseek.R1.32b.qwen.distill.q4_K_M, Gemma3.27b.it.q4_K_M, and Gemma3n.e4b.it.fp16. Our results show that the models annotate foreign commenters with an accuracy ranging from 76% to 95%. However, Cohen's Kappa for inter-rater agreement with the human annotation remains under 80%.'Thinking' models such as Qwen and Deepsek generally perform best, while the Gemma models have the lowest kappa.

We are currently running additional experiments on different varieties of English, including a grammar annotation task to identify the quotative *be like* and a semantic task annotating various levels of animacy in subject and object positions with the aim of reporting on the potential and limitations of small, locally-run open-weight LLMs for a range of corpus annotation tasks.

References

Baker, P., Schmück, H., & Qian, Y. (2025). *Automatic Image Tagging for Corpus Linguistics: A Multimodal Study of News Representations of Islam (1st ed.)*. Cambridge. https://doi.org/10.1017/9781009581233

Reiter, N. (2025). *Reproducibility when Working with Large Language Models: A Hallucination?* ReproducibiliTea in the HumaniTeas, University of Cologne. https://osf.io/5jnem

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, *29*(4), 534–561. https://doi.org/10.1075/ijcl.23087.yu

When no model fits – AI-based translation strategies for historical dialect corpora

Alfred Lameli

The paper addresses the research question of how current AI technologies – in particular large language models (LLMs) – can be applied to the translation of historical dialect texts. Historical dialect texts are of high linguistic relevance as they document language stages that no longer exist today and provide insights into the regional, social, and cultural contexts of past centuries. For their linguistic analysis, accurate translation and annotation are indispensable. However, this task is complicated by, for example, orthographic variation, lexical obsolescence, and the lack of digitally trained models for the dialects in question.

Against this background, the paper describes an ongoing project which aims to create a large, annotated corpus of 19th century dialects. At its core is a multi-stage translation pipeline that integrates AI-based translation methods with philological expertise. Translation into a standardized High German version serves as a bridge for automated processing steps and opens up new possibilities for annotation without erasing the specific features of the originals.

The paper explores the opportunities and limitations of working with LLMs in the context of extremely low-resource varieties. It discusses strategies for ensuring data quality and reproducibility, ways of integrating AI-assisted methods into the compilation of historical corpora, and perspectives for developing small, targeted language models for dialect processing.

Beyond the methodological framework, the project is presented as an interdisciplinary initiative combining linguistics, computer science, and digital humanities, with the long-term goal of making the resulting corpus freely available. This resource will not only support detailed linguistic analyses but also serve as a foundation for dialectometric and lexicographical work. The project thus contributes to the broader discussion on how AI technologies can be used in corpus linguistics to meet the specific requirements of historical dialect data – while maintaining the balance between automation and philological precision.