

VARVARA ARZT

📍 Vienna, Austria • 📞 +43 676 5880060 • ✉ varvara.arzt@tuwien.ac.at

🔍 Google Scholar • 🌐 kleines-gespenst • 📺 varvara-arzt • 🐦 @wienergespenst • kleines-gespenst.bsky.social

PhD researcher in NLP at TU Wien. I work on how language models acquire natural language, how that knowledge is stored, and whether different types of knowledge (e.g. factual and linguistic) are stored in different ways, including in multilingual settings that go beyond high-resource languages. My recent work used parallel experiments with controlled artificial and typologically similar natural languages to separate architectural biases from effects of the training data on LM word order preferences. **Research interests:** Mechanistic interpretability, and how far interpretability methods can themselves be trusted • Model diffing • LLM sycophancy • Multilingual NLP • Language modeling • AI safety

EDUCATION

PhD in Computer Science (in progress) September 2023 – expected autumn 2027
TU Wien, Vienna, Austria • Advisors: Terra Blevins and Allan Hanbury
Thesis: “Factual and Linguistic Knowledge in Language Models”

MA in Byzantine and Modern Greek Studies March 2017 – April 2022
University of Vienna, Vienna, Austria
Thesis: “Rebetiko Meets Digital Humanities: Automatic Analysis of Rebetiko Lyrics via Data Science Methods”

BA in Linguistics (focus: Albanology) September 2010 – June 2014
Saint Petersburg State University, Russia

RESEARCH POSITIONS

University Assistant (Prae-Doc), Data Science Research Unit September 2023 – present
TU Wien, Vienna, Austria

University Assistant (Prae-Doc), Digital Age Research Center (D!ARC) August 2022 – November 2025
University of Klagenfurt, Austria

Research Intern, Austrian Academy of Sciences (ACDH-CH) July – August 2021
Vienna, Austria • NLP and machine learning

SELECTED RESEARCH

- **Word order preferences in LMs are data-driven, not architectural.** Compared decoder-only LMs across 192 artificial languages and typologically diverse natural languages. The same architecture shows a left-branching preference on artificial languages but an emerging right-branching SVO advantage on natural ones as training data scales, and that advantage tracks resource level and data quality rather than word order itself, implying LLM adoption may gradually narrow word order diversity. (EMNLP 2026, main)
- **RE models pattern-match rather than generalise.** Cross-dataset experiments showing relation extraction models fail to transfer even within similar domains, and that higher in-dataset performance often signals overfitting to dataset artefacts rather than better transferability. Data quality, not lexical similarity, drives robust transfer, and zero-shot baselines occasionally beat every cross-dataset result. (IJCNLP–AAACL 2025, main)
- **RE benchmarks and leaderboards obscure real progress.** Audit of widely used RE benchmarks (TACRED, NYT) documenting severe class imbalance, noisy labels and undocumented construction decisions, and showing that aggregate F1 leaderboards hide per-class failure with recommendations applicable well beyond RE. (GenBench @ EMNLP 2024)

PUBLICATIONS

Peer-reviewed

- **Arzt, Varvara**, Allan Hanbury, and Terra Blevins. Left-Branching Transformers Excel at Right-Branching Languages: Data Shapes Word Order Preferences in Language Models. To appear in *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2026), main conference. Association for Computational Linguistics, 2026.

- **Arzt, Varvara**, Allan Hanbury, Michael Wiegand, Gábor Recski, and Terra Blevins. Relation Extraction or Pattern Matching? Unravelling the Generalisation Limits of Language Models for Biographical RE. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the ACL (IJCNLP–AACL 2025)*, pages 1463–1484, Mumbai, India. Association for Computational Linguistics.
- **Arzt, Varvara**, and Allan Hanbury. Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 120–130, Miami, Florida, USA. Association for Computational Linguistics, 2024.
- **Arzt, Varvara**, Mohammad Mahdi Azarbeik, Ilya Lasy, Tilman Kerl, and Gábor Recski. TU Wien at SemEval-2024 Task 6: Unifying Model-Agnostic and Model-Aware Techniques for Hallucination Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1183–1196, Mexico City, Mexico. Association for Computational Linguistics, 2024.
- Annisa Maulida Ningtyas, Guntur Budi Herwanto, Yunita Sari, Rifki Afina Putri, Filip Kovacevic, Alaa El-Ebshihy, **Varvara Arzt**, and Florina Piroi. AnnoHID: LLM-Assisted Annotation Framework for Low-Resource Medical Texts. In *Proceedings of the 64th Annual Meeting of the ACL (Volume 3: System Demonstrations)*, pages 683–691, San Diego, California, USA. Association for Computational Linguistics, 2026.

Preprints

- Tyler A. Chang, Catherine Arnett, et al. (350+ contributors, including **Varvara Arzt**). Global PIQA: Evaluating Commonsense Reasoning Across 100+ Languages and Cultures. arXiv:2510.24081, 2025. Community benchmark associated with the 5th Multilingual Representation Learning (MRL) workshop; contributed the Albanian subset.

Talks and posters

- **Varvara Arzt**. Analysis of Word Order Biases in Language Models: A Controlled Investigation Using Artificial and Natural Languages. 49th Austrian Linguistics Conference (Österreichische Linguistiktagung, ÖLT 2025), University of Klagenfurt, Austria, December 2025.
- Sandy Aoun, **Varvara Arzt**, Daniel Luger, and Georg Vogeler. Information Extraction from German Medieval Charters Abstracts. 19th International Conference of the Alliance of Digital Humanities Organizations (DH 2024), Washington, D.C., USA, August 2024.
- **Varvara Arzt**. Exploring the Narratives of Rebetiko via Corpus-Based Analysis. Storytelling, DARIAH Annual Event 2022, Athens, Greece, May 2022.

Earlier work in linguistics (published as V. A. Diveeva; name changed to Arzt in June 2019)

- Diveeva, V. A. Markirovaniye aktantov dvuchmestnich predikativ v albanskom yazike [Actant marking of two-place predicates in Albanian]. In S. S. Say (ed.), *Valentniye klassi dvuchmestnich predikativ v raznostruktturnich yazikach* [Valency classes of two-place predicates], 72–88. ILS RAS, Saint Petersburg, Russia, 2018.
- Four conference talks on Albanian argument marking, labile verbs and experiencer constructions. International Conferences for Philology Students, Saint Petersburg State University (2012–2014), and the Institute of Linguistic Studies, Russian Academy of Sciences (2014).

TEACHING

TU Wien and University of Klagenfurt, since Winter Term 2023:

- Natural Language Processing and Information Extraction (2023–2026)
- Information Retrieval (2024–2026)
- Data-oriented Programming Paradigms (2023, 2024)
- Programming in Python (2024)

RESEARCH PROJECTS

- **AnnoHID** (TU Wien / Universitas Gadjah Mada, Indonesia), 2025–2026. LLM-assisted annotation of biomedical text in low-resource languages: named entity recognition and medical concept normalisation. Demo published at ACL 2026.
- **GlagoLab** (TU Wien / University of Zadar), 2024–present. Information extraction from Glagolitic Croatian texts.
- **Albanian National Corpus**, ILS RAS, 2012–2014. Text collection and processing.
- **Non-canonical actant marking** (RSCI, PI: S. S. Say), ILS RAS, 2012–2014.
- **Contact zones of Russia, Ukraine, Moldova and Romania** (RSCI grant 12-31-01026, PI: A. A. Novik), MAE RAS, 2012–2014.
- **Field work**: Albanian villages of Ukraine, pronominal system of the Ukrainian Albanian dialect (summers 2011, 2012); southeastern Albania (September 2012).

ACADEMIC HONOURS

- Russian Presidential Scholarship for outstanding academic performance of young researchers, 2012 and 2013.

TECHNICAL SKILLS

- **Languages and tooling:** Python (PyTorch, HuggingFace Transformers, NumPy, SciPy, pandas, scikit-learn, spaCy, NLTK, Flair, Matplotlib), Git, L^AT_EX, Bash, Linux.
- **Model training and analysis:** training and fine-tuning transformer LMs; zero-shot and few-shot ICL; statistical analysis; MLflow and Weights & Biases; SLURM clusters and multi-GPU training.
- **Interpretability:** TransformerLens, nnsight, SAELens, Neuronpedia; sparse autoencoders, activation patching, probing.
- **Data:** corpus construction, annotation workflow design (Prodigy, Label Studio), inter-annotator agreement, web scraping (BeautifulSoup, Scrapy, Selenium).
- **Natural languages:** Russian (native), German (C2), English (C1), Albanian (C1), Modern Greek (C1), French (B1), Spanish (A2), Turkish (A1), Latin and Ancient Greek.

SUMMER SCHOOLS AND EXCHANGES

- BAL-ADRIA Summer School on Digital Humanities, University of Zadar, Croatia, June 2022.
- Helsinki Digital Humanities Hackathon, University of Helsinki, May 2022 (poster).
- Exchange semester, Byzantine and Modern Greek Studies, University of Vienna, 2013–2014; summer schools of Modern Greek (Athens 2013, Cyprus 2012) and Albanian (Tirana 2011, 2012).

EARLIER PROFESSIONAL EXPERIENCE

- Freelance interpreter and translator (Albanian, English, Greek, German, Russian), 2011–2024, including volunteer work for Caritas Kärnten (2023–2024) and secretary to the Honorary Consul of Albania in Saint Petersburg (2012–2014).
- Data Management Department, Oberösterreichische Lehrer- Kranken- und Unfallfürsorge (LKUF), Linz, Austria, 2018–2019.
- English teaching and tutoring, Saint Petersburg, 2014–2015.

OTHER

- Crew member of the frigate *Shtandart*, a replica of a ship from the time of Peter the Great, 2006–present; formerly pointswoman and engine driver on the Malaya October Railway, 2002–2007.
- Piano and cello, sailing, running, literature.