Stephanie Evert X **Research** X Teaching X CV X Publications X Software X Hobbies X Trans*

Research Interests

My **computational corpus linguistics group** at FAU Erlangen–Nürnberg <http://www.linguistik.fau.de/> carries out foundational methodological research on the quantitative analysis of large text corpora. The algorithms and software tools developed by the group support innovative studies in the digital humanities and social sciences as well as practical applications in language technology. A particular focus lies on understanding cooccurrence phenomena and their application in corpus-based discourse analysis.

Methodological foundations X Corpus tools X Cooccurrence phenomena

Methodological foundations of corpus research and digital humanities

Corpus research in linguistics as well as in the digital humanities and social sciences relies on a wide range of statistical techniques and visualizations. A central goal of my research is to develop sound methodological foundations for corpus linguistics, which address key problems in order to ensure that quantitative analyses are both reliable and meaningful.

Projects

- 2014X2019: **Kallimachos** <https://www.linguistik.phil.fau.de/projects/kallimachos/> (BMBF e–Humanities-Zentrum led by U Würzburg <http://www.kallimachos.de/>)

  The FAU sub-project was concerned with methodological issues and the interpretation of quantitative measures in literary stylometry, focussing on authorship attribution (phase 1) and lexical/syntactic complexity (phase 2).

Software

- **zipfR**: R package for LNRE modelling of type-token distributions X <http://zipfR.r–forge.r–project.org/>

Key publications

- Evert et al. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*. **22**(suppl_2), ii4Xii16. [free access (PDF) <https://doi.org/10.1093/llc/fqx023>, reference corpus <https://github.com/cophi-wue/refcor>].
- Evert & Neumann (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In: *Empirical Translation Studies. New Theoretical and Methodological Traditions*, TiLSM number 300. [online supplement <http://www.stefan-evert.de/PUB/EvertNeumann2017/>]
- Schäfer at al. (2017). Japan's 2014 general election: Political bots, right-wing internet activism and PM Abe ShinzX's hidden nationalist agenda. *Big Data*, **5**(4), 294X309. [open access (PDF) <https://doi.org/10.1089/big.2017.0049>]
- Evert et al. (2017). Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch. In: *Proceedings of Corpus Linguistics 2017*. [PDF <http://purl.org/stefan.evert/PUB/EvertWankerlNoeth2017.pdf>]
- Evert & Arppe (2015). Some theoretical and experimental observations on naïve discriminative learning. In: *Proceedings of QITL–6*. [PDF <http://purl.org/stefan.evert/PUB/EvertArppe2015.pdf>]
- Proisl etc. (2018). Delta vs. n–gram tracing: Evaluating the robustness of authorship attribution methods.. In *Proceedings of LREC 2018*. [PDF <http://purl.org/stefan.evert/PUB/ProislEtc2018_LREC.pdf>, slides <http://purl.org/stefan.evert/PUB/ProislEtc2018_LREC_slides.pdf>]
- Baroni & Evert (2007). Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In: *Proceedings of ACL 2007*. [PDF <http://purl.org/stefan.evert/PUB/BaroniEvert2007.pdf>]
- Evert (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* **54**(2). [manuscript (PDF) <http://purl.org/stefan.evert/PUB/Evert2006.pdf>]

Corpus tools and language technology

My group develops algorithms and software tools for the automatic linguistic annotation, efficient indexing, flexible query and quantitative analysis of large text corpora. These tools form the basis of innovative research in the digital humanities as well as practical and commercial applications in language technology.

Projects

- 2021X2023: **RAND** <https://cris.fau.de/converis/portal/Project/248596472?lang=en_GB> X Reconstructing Arguments from Newsworthy Debates (DFG SPP 1999: RATIO)

  2018X2020: **RANT** <https://cris.fau.de/converis/portal/Project/125900111?lang=en_GB> X Reconstructing Arguments from Noisy Text (DFG SPP 1999: RATIO)

  A corpus-linguistic approach to argumentation mining in social media, combined with representation and inference in a powerful logical framework.

- 2020X2022: **LeAK** <https://cris.fau.de/converis/portal/Project/264601690?lang=en_GB> X Automatic anonymisation of German court decisions (research contract from BayStMJ <https://www.justiz.bayern.de/>)

  This project explores the feasibility of fully automatic anonymisation of German court decisions. Its key contributions are the creation of a high-quality manually annotated gold standard and the thorough evaluation of automatic algorithms.

Software & resources

- **CWB**, the IMS Open Corpus Workbench for indexing & querying large text corpora X <http://cwb.sf.net/>
- **EmpiriST corpus**, a gold standard for linguistic annotation of German Web & CMC texts X <https://github.com/fau−klue/empirist−corpus/>
- **Web1T5−Easy** indexes Google Web n−grams with SQLite X http://webascorpus.sf.net/ <http://webascorpus.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_10_Software&subpage=FILES_50_Goog Grams>

Key publications

- Evert & Hardie (2015). Ziggurat: A new data model and indexing format for large annotated text corpora. In: *Proceedings of CMLC−3*. [PDF <http://purl.org/stefan.evert/PUB/EvertHardie2015.pdf>]
- Evert et al. (2020). Corpus Query Lingua Franca part II: Ontology. In *Proceedings of LREC 2020*. [PDF <http://purl.org/stefan.evert/PUB/EvertEtc2020_CQLF2.pdf>, GitHub <https://github.com/cqlf-ontology/cqlf>]
- Evert et al. (2016). A distributional approach to open questions in market research. *Computers in Industry* **78**. [manuscript (PDF) <http://purl.org/stefan.evert/PUB/EvertGreinerEtc2016_COMIND.pdf>]
- Proisl et al. (2020). EmpiriST corpus 2.0: Adding manual normalization, lemmatization and semantic tagging to a German Web and CMC corpus. In *Proceedings of LREC 2020*. [PDF <http://purl.org/stefan.evert/PUB/ProislEtc2020_LREC.pdf>, corpus & resources <https://github.com/fau-klue/empirist-corpus>]
- Evert et al. (2014). SentiKLUE: Updating a polarity classifier in 48 hours. In: *Proceedings of SemEval 2014*. [PDF <http://purl.org/stefan.evert/PUB/EvertEtc2014_SentiKLUE.pdf>]
- Evert & Hardie (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In: *Proceedings of Corpus Linguistics 2011*. [PDF <http://purl.org/stefan.evert/PUB/EvertHardie2011.pdf>]
- Evert (2010). Google Web 1T5 n−grams made easy (but not for the computer). In: *Proceedings of WAC−6*. [PDF <http://purl.org/stefan.evert/PUB/Evert2010_WAC6.pdf>, Web1T5−Easy]
- Giesbrecht & Evert (2009). Part-of-speech tagging − a solved task? An evaluation of POS taggers for the Web as corpus. In: *Proceedings of WAC−5*. [PDF

<http://purl.org/stefan.evert/PUB/GiesbrechtEvert2009_Tagging.pdf>]

Collocations, multiword expressions and corpus-based discourse analysis

Cooccurrence patterns X such as collocations, multiword expression, valency and distributional semantics X play a central role not only in corpus linguistics but also for studying public discourses and political propaganda. My research in this area focuses on improving and refining the underlying analytical techniques as well as the development of new interactive methods for multi-modal corpus-based discourse analysis.

Projects

- 2022X2024: **NormRechts** X The Normalization of Right-wing Populist and New Right Discourses in Japan and Germany (DFG)

  This project will further develop the MMDA methodology and apply it to the comparative analysis of right-wing populist discourses in Japan and Germany.

- 2021X2022: **Tracking the Infodemic** <https://cris.fau.de/converis/portal/Project/252040566?lang=en_GB>: Conspiracy theories in the corona crisis (VolkswagenStiftung)

  This research project applies innovative corpus-linguistic methods to analyse the use and distribution of typical linguistic patterns of conspiracy theories and study the discursive strategies they share with right-wing populist and extremist discourses.

- 2017X2019: **Exploring the Fukushima Effect** <https://www.linguistik.phil.fau.de/projects/efe/> (FAU Emerging Fields Initiative)

  XAttitudes and Opinions towards Nuclear Power and Renewable Energy and the Emergence of a Transnational Algorithmic Public Sphere.X A key contribution of this project is the development of the innovative MMDA methodology and software toolkit for corpus-assisted discourse analysis.

Software

- **MMDA** <https://www.linguistik.phil.fau.de/projects/efe/mmda-toolkit/>: an interactive software tool for corpus-assisted discourse analysis
- The **UCS Toolkit** for collocation research X <http://www.collocations.de/software.html>
- **wordspace**: an R package for distributional semantics X <http://wordspace.r−forge.r−project.org/>

Key publications

- Evert (2008). Corpora and collocations. In: *Corpus Linguistics. An International Handbook.* [extended manuscript (PDF) <http://purl.org/stefan.evert/PUB/Evert2007HSK_extended_manuscript.pdf>]
- Lapesa & Evert (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics* **2**. [PDF <http://purl.org/stefan.evert/PUB/LapesaEvert2014tacl.pdf>, supplementary material <http://www.linguistik.fau.de/dsmeval/>]
- Heinrich et al. (2018). A transnational analysis of news and tweets about Xnuclear phase-outX in the aftermath of the Fukushima incident. In *Proceedings of the CIDTD 2018 Workshop*. [PDF <http://purl.org/stefan.evert/PUB/HeinrichEtc2018_CIDTD.pdf>, slides <http://purl.org/stefan.evert/PUB/HeinrichEtc2018_CIDTD_slides.pdf>]
- Evert (2014). Distributional semantics in R with the wordspace package. In: *Proceedings of COLING 2014*. [PDF <http://purl.org/stefan.evert/PUB/Evert2014_wordspace.pdf>, wordspace homepage <http://wordspace.r-forge.r-project.org/>]
- Evert et al. (2017). E−VIEW-alation X a large-scale evaluation study of association measures for collocation identification. In *Proceedings of eLex 2017*. [PDF <http://purl.org/stefan.evert/PUB/EvertUhrigEtc2017.pdf>, slides <http://purl.org/stefan.evert/PUB/EvertUhrigEtc2017_slides.pdf>, video <https://www.youtube.com/watch?v=xYo3wTRx8F8>, E−VIEW-alation

<http://www.collocations.de/eviewalation/>]

- Evert & Krenn (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language* **19**(4). [manuscript (PDF) <http://purl.org/stefan.evert/PUB/EvertKrenn2005.pdf>]