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

- **Kallimachos** (BMBF e–Humanities-Zentrum) <http://www.kallimachos.de/>

Software

- **zipfR**: R package for LNRE modelling <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*, advance access. [PDF <https://academic.oup.com/dsh/article/doi/10.1093/llc/fqx023/3865676/Understanding-and-explaining-Delta-measures-for?guestAccessKey=54b7daa4-be40-4687-880a-543d2b41254e>]
- 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/>]
- 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>]
- Diwersy et al. (2014). A weakly supervised multivariate approach to the study of language variation. In: *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. [manuscript (PDF) <http://purl.org/stefan.evert/PUB/DiwersyEvertNeumann2012.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.

Software

- **CWB**, the IMS Open Corpus Workbench for indexing & querying large text corpora <http://cwb.sf.net/>

- **Web1T5−Easy** indexes Google Web n−grams with SQLite 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. (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>]
- Evert (2016). CogALex-V shared task: Mach5 X a traditional DSM approach to semantic relatedness. In: *Proceedings of CogALex-V*. [PDF <http://purl.org/stefan.evert/PUB/Evert2016_Mach5.pdf>, system & data <http://www.collocations.de/data/#mach5>]
- Bei*ß*wenger et al. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In: *Proceedings of WAC-X*. [PDF <http://purl.org/stefan.evert/PUB/BeisswengerEtc2016.pdf>, task homepage <https://sites.google.com/site/empirist2015/>]
- 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

- **Exploring the Fukushima Effect** (FAU Emerging Fields Initiative)

  Attitudes and Opinions towards Nuclear Power and Renewable Energy and the Emergence of a Transnational Algorithmic Public Sphere

Software

- **UCS Toolkit** for collocation research <http://www.collocations.de/software.html>
- **wordspace**: an R package for distributional semantics <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/>]

- 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/>]
- Michelbacher et al. (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* **7**.
- 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>]