Contribution ID: **59** Type: **Poster**

Linguistic corpus research software at the Leibniz-Institute for the German Language (IDS)

Wednesday 26 February 2025 19:40 (20 minutes)

Research in linguistics is increasingly data-driven and requires access to language corpora, i.e. "collection[s] of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language"(Crystal 2003). Here, language itself is the object of study, and not just an obstacle on the way to retrieve information.
Building large corpora for scientifically valid research is a labour-intensive process. This is true for written language, and even more so for spoken language, which not only needs to be converted into written form, but which also contains multiple, potentially overlapping speakers. Corpora also have to be enriched with relevant meta-information and linguistic annotations. What is more, the data that goes into a corpus can be subject to copyright (e.g. newspapers) or personal/privacy rights limitations (e.g. recorded and transcribed private conversation), which need to be sorted out before the corpus can be compiled and used. Thus, high-quality language corpora and the research software for querying, exploring, analysing and visualizing them are valuable assets for the linguistic research community. The Leibniz-Institute for the German Language (IDS) provides collections of both written and spoken language corpora and the specialized corpus research software for accessing them.
Corpus-based linguistic research ranges from simple corpus search for retrieving instances of certain language phenomena to large-scale training of language models, whereby a concurrent reference to metadata (external) and content data (internal) is possible (Sinclair 1996). Other research tasks include the creation of statistics about the frequency of words or word combinations, or quantitative analyses of more complex linguistic structures. In order to be scientifically valid, the respective results need to be reproducible, ideally over a longer time.
Belonging to the domain of humanities, linguistics has a higher share of practitioners with little technical literacy, which imposes limits on how difficult the use of the corpus research software should be. On the other hand, some advanced research questions simply require more powerful and thus more technically demanding methods, which many researchers, in particular from computational linguistics, actually have.
At the IDS, corpus creation and software development and operation take place in projects with permanent funding, which ensures long-term availability. Access for registered users is only provided via web UIs or APIs, thus protecting the integrity of the data. Our aim is to provide as large a database as possible from which users can compile sub-corpora according to their research question by applying meta-data criteria. Using the exact same data basis, the software offers access via easy-to-use form-based query templates or graphical assistants, but also via specialized corpus query languages for advanced users. While the corpora are continuously expanded, changes between versions are tracked in the corpus meta-data, allowing to reproduce results from earlier versions as required.
Our poster outlines how the IDS approaches the various conceptual, legal, linguistic, and technical challenges of research software for written and spoken corpora.

I want to participate in the youngRSE prize

no

Primary authors: Dr MÜLLER, Mark-Christoph (Leibniz-Institut für deutsche Sprache); DIEWALD, Nils (Leibniz-Institut für deutsche Sprache); FRICK, Elena (Leibniz-Institut für deutsche Sprache); Dr KUPIETZ, Marc (Leibniz-Institut für deutsche Sprache)

Presenter: Dr MÜLLER, Mark-Christoph (Leibniz-Institut für deutsche Sprache)

Session Classification: Poster and Demo Session together with Reception

Track Classification: Research Software: digital humanities