Alexander Geyken, Marc Kupietz

Editorial

With the growing availability and importance of (large) corpora in all fields of linguistics, the role of software tools is gradually moving from useful, possibly intelligent information-technological "helpers" towards scientific instruments that are as integral parts of the research process as data, methodology and interpretations. Both aspects are present in this special issue of JLCL on *corpus linguistic software tools*: The contributions address topics such as software tools for managing corpora, transforming corpus data, annotating and analyzing corpora as well as innovative ways of exploring and visualizing corpus data and analyses.

Thomas Krause, *Ulf Leser* and *Anke Lüdeling* present graphANNIS, a newly developed graph database for querying deeply annotated linguistic corpora. After discussing the pros and cons of existing solutions, they document implementation details, the data model, and query aspects, including the evaluation of the the proposed graph based implementation vs. the established RDBMS backend. It is reported that graphANNIS performs in the majority of the tested cases much faster than the relational equivalent.

Paul Rayson, *John Mariani*, *Bryce Anderson-Cooper*, *Alistair Baron*, *David Gullick*, *Andrew Moore* and *Stephen Wattam* address in their paper "towards interactive multidimensional visualisations for corpus linguistics" several important issues, including how iterative and exploratory corpus investigations can be supported by dynamic and interactive visualisation techniques and how to approach issues with current corpus retrieval tools by including interactive and multidimensional visualizations.

Christian Pölitz presents in his paper "data mining software for corpus linguistics with an application in diachronic linguistics" a freely available plugin for the RapidMiner software. This plug-in could be of particular relevance for researchers already using RapidMiner, or for computational linguists looking for a tool that is more user-friendly - through drag-and-drop - than e.g. the established R software. The author uses the example case of topic modelling over time to demonstrate the advantages of the new plugin, including the implemented topic models and coherence measures.

Nils Diewald and *Eliza Margaretha* present "Krill", the search component of the KorAP corpus search and analysis platform. This paper makes three important contributions to research and software engineering in the area of corpus indexing and query: It introduces a new open-source corpus indexing software based on Apache Lucene and describes how linguistic corpus search can be implemented on top of a full text search engine such as Lucene. Furthermore, as an implementation of the KoralQuery specification, Krill is an important milestone for the types of search pattern expected from a modern linguistic corpus query engine.

Maria Skeppstedt, *Carita Paradis* and *Andreas Kerren* present PAL, a standalone tool for pre-annotation and active learning. PAL can be trained using gold standard data

and it can incorporate manually annotated/corrected data created by the user within the BRAT annotation tool. PAL tries to optimize the process by using an active learning approach to select sentences to be annotated by the user. The tool focuses on "chunk"-based annotation tasks, e.g. for named entity detection.

Arne Neumann presents “discourse graphs”, a library and command-line application for merging and validating heterogeneous, multi-layered corpora. The resulting software tool consists of a python library and command-line applications for converting multi-level annotated data into different formats, merging independent annotations of the same text into one graph representation and validating heterogeneous annotation layers of the same text. This complex model is built upon the property graph model and can therefore benefit from the ecosystem built around modern graph libraries based on that paradigm.

Thomas Schmidt presents software tools and workflows for the construction and dissemination the FOLK corpus, a corpus of spoken interaction. The article covers the tools used in the individual steps of transcription, anonymization, orthographic normalization, lemmatization and POS tagging of the data, as well as the utilities used for corpus management. Furthermore, it presents the DGD (Datenbank für Gesprochenes Deutsch - Database of Spoken German) as means for distributing the completed research data and making it available for qualitative and quantitative analysis.

Ruprecht von Waldenfels and *Michał Woźniak* present SpoCo, an adaptable query and analysis system for spoken dialect corpora with aligned audio data encoded in ELAN. SpoCo is targeted at users of different levels of expertise and provides them with advanced concordancing functions, as well as the the possibility to edit and correct transcriptions. SpoCo takes a middle position between systems that are developed for the purposes of a specific dialect corpus, on the one hand, and general-use systems designed for a wide range of data and usage cases, on the other. It is used in a network of Slavic dialect projects that cooperate in tool development and data sharing.

Our gratitude goes to the colleagues who contributed to this special issue as external reviewers.

Berlin and Mannheim, May 2017

Alexander Geyken
Marc Kupietz