KorAP: The New Corpus Analysis Platform at IDS Mannheim

Piotr Bański[1], Joachim Bingel[1], Nils Diewald[1], Elena Frick[1], Michael Hanl[1], Marc Kupietz[1], Piotr Pęzik[2], Carsten Schnober[1], Andreas Witt[1]

[1]Institut für Deutsche Sprache, Mannheim
{banski,bingel,diewald,frick,hanl,kupietz,schnober,witt}@ids-mannheim.de

[2]University of Łódź
piotr.pezik@gmail.com

Abstract

The KorAP project ("Korpusanalyseplattform der nächste Generation", "Corpus-analysis platform of the next generation"), carried out at the Institut für Deutsche Sprache (IDS) in Mannheim, Germany, has as its goal the development of a modern, state-of-the-art corpus-analysis platform, capable of handling very large corpora and opening the perspectives for innovative linguistic research. The platform will facilitate new linguistic findings by making it possible to manage and analyse extremely large amounts of primary data and annotations, while at the same time allowing an undistorted view of the primary un-annotated text, and thus fully satisfying expectations associated with a scientific tool. The project started in July 2011 and is funded till June 2014. The demo presentation in December will be the first version following a preliminary feature freeze, and will open the alpha testing phase of the project.

Keywords: corpus analysis, written corpora, scalability, virtual collections, CQLF, standoff annotation, search engine

1. Introduction

The KorAP project ("Korpusanalyseplattform der nächsten Generation", "Corpus-analysis platform of the next generation"), carried out at the Institut für Deutsche Sprache (IDS) in Mannheim, Germany, has as its goal the development of a modern, state-of-the-art corpus-analysis platform, capable of handling very large corpora and opening the perspectives for innovative linguistic research. In its design, KorAP follows the well-established paradigm of an open client-server architecture with a modular structure, strictly separating the backend part from the frontend. However, we have decided to leverage the richness of the current proposals for various kinds of information retrieval and to employ two backends via a single interface. One is based strictly on the Lucene platform[1], and the other employs Lucene as a filter that reduces the search space in a parallelized graph database, Neo4j[2] (see fig. 1). These backends can be switched around, but we have also been experimenting with query mapping designed to take advantage of both of them at the same time.

Features of KorAP that we are particularly willing to present at the Conference include the following:

- support for virtual collections,
- support for multiple, potentially conflicting, annotations,
- ability to prevent the imposition of the theoretical views of the annotator upon the user,
- sophisticated display abilities,
- initial reference implementation of the ISO TC37 SC4 WG6 Work Item "Corpus Query *Lingua Franca*".

In what follows, we briefly describe some of the above-mentioned features, while the others, due to the space limitations, shall be presented and discussed during the demonstration session.

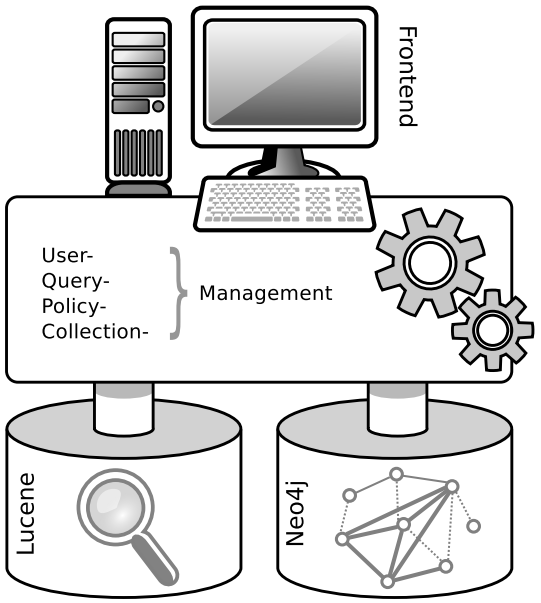

Figure 1: The KorAP Retrieval System Architecture

2. KorAP at a glance

KorAP implements a theory-neutral approach to modelling linguistic resources, by adopting the radical stand-off architecture for the document data model. This means keeping the base text as a single sequence of characters and placing the (possibly conflicting) linguistic analyses in separate compartments (which we refer to as "foundries", cf. fig. 2).

The user can query a single foundry that she recognizes as being appropriate for the given task, or can even

[1]https://lucene.apache.org/
[2]http://www.neo4j.org/

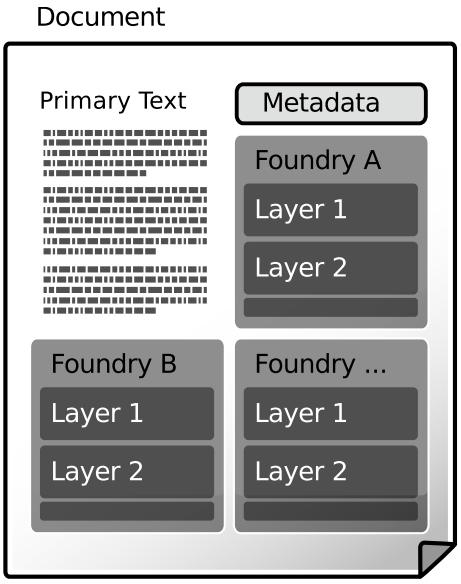

Figure 2: The KorAP Document Model

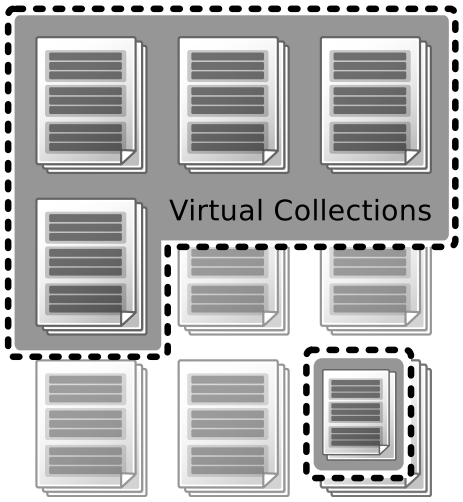

Figure 3: Virtual collections can be defined as subsets or supersets of corpora

query multiple foundries and multiple linguistic layers at the same time, benefitting from the strengths of different sources and exploiting complementary information expressed in parallel annotations. As an example, dependency and constituency annotations of 'syntactic words' can be cross-linked with more fine-grained part-of-speech categories that are available for the corresponding spans of text defined inside foundries with morphosyntactic annotations. Such links can be used directly in the Neo4j backend as special filtering conditions in corpus queries.

Note that by placing all annotation information inside separate foundries and by keeping the primary text as a single sequence of characters, we are not privileging any particular point of view on the underlying data (text), and in this way we ensure that the analysis is not biased. Naturally, this approach carries an increased cost in terms of maintenance of the connection between the individual annotation layers and the particular spans that these annotations describe. However, because the logical separation of documents carries over into the index structure, modifications of this type are always localized, with only the immediately affected annotation layers requiring adjustments, and with only specific parts of the Lucene index getting recreated.

KorAP makes it possible to assemble virtual collections out of possibly opportunistic corpora, on the basis of various text-internal (e.g., the presence of a desired wordform) and text-external (author, date of publication, etc.) parameters, cf. Bański et al. (2013), see fig. 3.

KorAP is designed to serve as a reference implementation of the new ISO TC36 SC4 proposal "Corpus Query Lingua Franca" (CQLF, cf. Mueller, 2010). In our system, CQLF occupies the space between query language parsers and the query builders for both backends, and serves as the matrix, onto which various particular QL constructs are mapped and from which they are serialized and uniformly carried over into the backend part. Currently, apart from the internal query language used for testing and debugging, KorAP has support for two established corpus query languages, COSMAS II (Bodmer, 2005) and Poliqarp (Janus and Przepiórkowski, 2007).

KorAP is able to visualize syntactic trees or dependency graphs as well as overlapping spans in a complex KWIC display. The project started in July 2011 and is funded till June 2014. The demo presentation in December will be the first version following a preliminary feature freeze, and will open the alpha testing phase of the project.

References

Bański, P., E. Frick, M. Hanl, M. Kupietz, C. Schnober, and A. Witt, 2013. Robust corpus architecture: a new look at virtual collections and data access. In A. Hardie and R. Love (eds.), *Corpus Linguistics 2013 Abstract Book*. Lancaster: UCREL. `http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf`.

Bodmer, F., 2005. COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport*, 3/2005:2–5.

Janus, D. and A. Przepiórkowski, 2007. Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics.

Mueller, M, 2010. Towards a digital carrel: A report about corpus query tools. Technical report. Unpublished report submitted to the Mellon Foundation. `http://panini.northwestern.edu/mmueller/corpusquerytools.pdf`.