Published in: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Doğan, Mehmet Uğur/Maegaard, Bente/Mariani, Joseph/ Moreno, Asuncion/Odijk, Jan/ Piperidis, Stelios: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey, May 2012. - European Language Resources Association (ELRA), 2012. pp. 2905-2911.

# The New IDS Corpus Analysis Platform: Challenges and Prospects

**Piotr Bański*, Peter M. Fischer, Elena Frick, Erik Ketzan,**
**Marc Kupietz, Carsten Schnober, Oliver Schonefeld, Andreas Witt**

Institute for the German Language (IDS)
R5 6–13, 68161 Mannheim, Germany
{banski|fischer|frick|ketzan|kupietz|schnober|schonefeld|witt}@ids-mannheim.de

**Abstract**

The present article describes the first stage of the KorAP project, launched recently at the Institut für Deutsche Sprache (IDS) in Mannheim, Germany. The aim of this project is to develop an innovative corpus analysis platform to tackle the increasing demands of modern linguistic research. The platform will facilitate new linguistic findings by making it possible to manage and analyse primary data and annotations in the petabyte range, while at the same time allowing an undistorted view of the primary linguistic data, and thus fully satisfying the demands of a scientific tool. An additional important aim of the project is to make corpus data as openly accessible as possible in light of unavoidable legal restrictions, for instance through support for distributed virtual corpora, user-defined annotations and adaptable user interfaces, as well as interfaces and sandboxes for user-supplied analysis applications. We discuss our motivation for undertaking this endeavour and the challenges that face it. Next, we outline our software implementation plan and describe development to-date.

**Keywords:** Corpus Linguistics, Corpus Analysis, Very Large Corpora, German Reference Corpus, KorAP

## 1. Introduction

Systematically assembled collections of communication acts, known as corpora, are now the most important empirical foundation of the field of linguistics (Lüdeling & Kytö, 2008). Corpora are used to confirm or refute hypotheses and constitute the primary basis of exploratory linguistic research (Tognini-Bonelli, 2001). To make large corpora manageable, the right tools are necessary to handle large volumes of data and perform computer-intensive analyses.

The Archiv für Gesprochenes Deutsch (AGD, "Archive of spoken German") (Fiehler et al, 2007) and the German Reference Corpus (DeReKo) (Kupietz & Keibel, 2009) comprise the world's largest collection of German-language data, and are stored at the Institut für Deutsche Sprache (IDS) in Mannheim, Germany. The Corpus Search, Management and Analysis System (COSMAS I, succeeded by COSMAS II) was created at the IDS to provide access to those corpora, and has been in operation since 1991 (cf. al-Wadi, 1994 and Bodmer, 2005).[1] COSMAS II was conceived in the early 1990s and implemented in the mid- and late 1990s. Providing nearly 20,000 users with access to a 5.4 billion word corpus, it is still one of the most powerful corpus-analysis platforms. Over time, however, COSMAS II has become increasingly difficult to update and modify in order to meet current and expected demands, such as, among others, the ability to manage corpora containing more than 10 billion words with multiple and potentially concurring annotation layers. This is also true of all the other corpus-analysis platforms that we know of.

The present contribution describes KorAP ("Korpusanalyseplattform der nächsten Generation"), a new project carried out at IDS Mannheim. The aim of this project is to develop a modern, state-of-the-art corpus-analysis platform, capable of handling very large corpora and opening the perspectives for innovative linguistic research. In what follows, we review the motivation for undertaking this endeavour and the challenges that face it. We also sketch the plans for the software implementation and describe development to-date

## 2. Challenges ahead

New trends in the field of linguistics have influenced research methods and changed the range of applications expected of new analytic tools. The success and proliferation of e-science within the Humanities (giving rise to the discipline now commonly referred to as e-Humanities) have been accompanied by an increased emphasis on empirical and scientific research. The importance of research data, along with such scientific maxims as falsifiability and reproducibility, has been growing in the humanities. Due to the challenges of large data volumes and the dynamic nature of many kinds of corpora, a strong need for traceability and reproducibility of linguistic research is becoming increasingly evident (cf. Pedersen 2008).

Linguists now also want to be able to collaboratively annotate and edit data, regardless of the location – this has led to the creation of research infrastructures such as CLARIN (Váradi et al, 2008) and virtual research environments such as TextGrid (Neuroth et al, 2011). This means that any system created from now on must have interfaces that communicate with such distributed infrastructures, and support functions and content such as federated search and analysis, re-usable distributed virtual corpora, and user-supplied annotations.

The immense growth of corpora in general, and DeReKo in particular, has raised new qualitative issues. The paradigm of data-driven analysis was, until recently, mostly relevant to the field of lexicology. But now the existence of very large text samples allows the analysis of complex linguistic

*IDS Mannheim and Institute of English Studies, University of Warsaw

[1] http://www.ids-mannheim.de/cosmas2/

structures and syntagmatic patterns, as well as their combinations with other factors such as time and origin (cf. Keibel et al. 2008). These new developments in grammar research and also linguistic theory as a whole are, for instance, reflected in conference series such as Grammar and Corpora (cf. e. g. Šticha & Fried, 2008) and new journals such as Corpus Linguistics and Linguistic Theory. As a result, corpus linguists are concerned with the creation of expensive and complex new research tools and, more generally, with finding support for new scientific research methods. These new methods include strategies that combine different approaches, for instance both data-driven and hypothesis-based, or different kinds of data, for instance both primary data and interpretive secondary data such as automatically generated linguistic annotations (cf. e. g. Müller, 2007 and Belica et al., 2011). The new corpus analysis platform developed at the IDS is projected to meet the above challenges, and to ultimately generalize beyond DeReKo, which is used as its primary testbed.

In addition to the above-mentioned issues, corpus linguistics is increasingly being confronted with the requirement to be able to handle data in different modalities. Multimodal resources, especially recorded and/or transcribed speech, are well-established sources of research data and require treatment according to established specifications. Although in the first phase of the project, we concentrate on written data, the new platform – by virtue of its modular and consequently multifunctional structure – is planned to be compatible with, and receptive to, multimodal data. In the first phase of the implementation, the ability to link multimodal data (facsimiles, audio, and video streams) to the text corpus will be supported. Such links will also point to and thereby virtually integrate specialised environments for research on spoken language developed in the context of the AGD. In subsequent stages of the platform development, we plan to integrate specialized modules for quantitative research on spoken language.

Challenges arise also in the field of computer science. Given the differences in budget, size and aims (information retrieval vs. linguistic research), search engine companies such as Google cannot be used as proof that dealing with data sets in the petabyte range is currently feasible for institutes and universities when aiming at linguistic research. That is because linguistic research should ideally remain accurate and reproducible[2] in order to satisfy scientific requirements (cf. Kilgarriff, 2007) and all the assumptions that underlie analyses should be transparent. Linguistics also introduces some seemingly insignificant issues that, nevertheless pose serious technical challenges, such as the requirement that common words or function words (articles, auxiliaries, etc.) cannot be ignored as they are in information retrieval.

Furthermore, a linguistic search must be able to query complex data structures and relationships, such as multiple metrics and levels, as well as to cope with the demanding requirements of the underlying inquiry (e. g. relations, quantifiers, regular expressions, etc.). Finally, the order of search results displayed must be controllable by users to allow them to select random samples.

The rights of third parties, especially copyright and privacy rights, almost always affect linguistic research data, and create challenges for the development of a corpus analysis system. Because these legal rights must be respected, access to corpora will typically be restricted by license terms and the contours of copyright law. The permissions that such licenses grant are, due to their high costs, typically very limited, especially regarding their transferability to the end user. An essential task in software development is therefore to make the data accessible to as large an extent as possible, while at the same time satisfying license conditions and rights holders' interests and using technical methods to prevent abuse.

Due to varying licensing conditions, as well as different user classes and use types, customizable security concepts have to be developed and implemented in order to make it possible to assign different rights to different user groups, and different restrictions to different resources. On the other hand, new ways to maximize the usefulness of the data have to be scouted. For instance, if data is bound to a location by license agreements, analysis and annotation software should nevertheless be able to access it. This goal can probably best be achieved by following Jim Gray's (2003, p 6) now famous postulate, "put the computation near the data". This means that instead of sending terabytes of copyright- or license-protected data through the Net, mobile-code sandboxes should be implemented. Unlike the primary data, analysis and annotation programs are typically created by the the research community and/or publicly funded and can be run locally in a controlled environment, without violating copyright law or license terms. As the context of Gray's claim suggests, such a strategy seems natural not only for the resolution of legal problems, but also from a computer science perspective for managing data-intensive problems in general and it also seems suitable with regard to the above-mentioned aspects of federated and distributed research. In any case, the grid metaphor seems much more appropriate for huge amounts of data than the cloud metaphor that for most resources is rendered impossible by license restrictions.

Along with the creation and implementation of the platform software, a top priority is the proper design of the data model and the storage, in order to guarantee a sufficiently efficient way to technically and practically manage huge amounts of data. The highest priority of the software development, however, is to secure an unbiased and undistorted view of the primary data. In keeping with Sinclair's (1994) Minimal Assumption Postulate we intend to keep the compromises that are necessary to tackle this task as minimal and as transparent as possible. This means, for example, not assuming a single possible tokenization or a single possible opinion on part-of-speech categories, even though this would make the software much more efficient. In this sense, also from the point of view of software implementation, the desired product is not merely a tool but rather a scientific tool. Transparency in this case also means that we intend to release as much of the software as possible (with the exception of, for instance, the heuristics for abuse detection)

[2] This means as reproducible as external conditions, including legal restrictions, allow. DeReKo, for example, receives requests for text withdrawal about once every week.

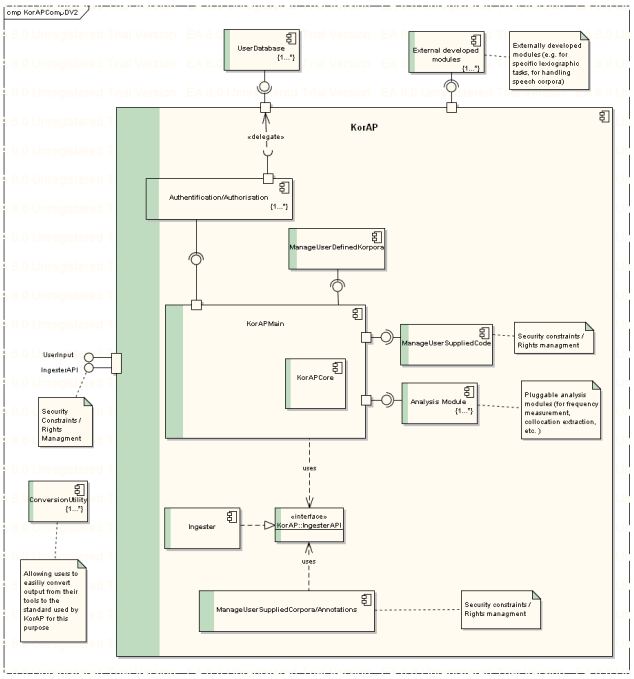


Figure 1: The modular KorAP architecture: general view

under an open-source license.

## 3. KorAP architecture: bird's eye view

The design of the KorAP platform follows the well-established paradigm of an open client-server architecture with a modular structure (Figure 1), strictly separating the back end from the various front ends, through which the user may interact with the system. Not only does this allow for flexibility and make it easy to extend the platform, it also complies with our iterative development approach: due to the uncertainty of whether all our goals are currently satisfiable, we want to be able to test modules separately, isolate the failing ones, and replace them or work around them. Future developments will be implementable in the form of new modules, so that partial requirements that we might fail to achieve today can still be achievable tomorrow. At the current stage of the project, the vast majority of the development (design and programming) focuses on the database – the core of the backend that includes the primary data, metrics, metadata and multiple concurrent annotation levels. The system has been designed in such a way as to be able to represent discontinuous structures and various sorts of relationships between the annotated objects. In order to improve scalability and system stability, architectures involving distributed databases and distributed file systems are being evaluated. The backend is being built with a view to supporting various adaptive or incremental indexing strategies, depending on data type. Another important component in the backend is the authentication and authorization framework that deals with user identity and the license status of the data. The corpus management components are designed to allow versioning and assignment of persistent identifiers (cf. Broeder et al., 2007), as well as dynamic creation of virtual corpora (cf. Kupietz et al., 2010). At a later stage, a web-based adaptive front end will be created, which the various user groups may interact with according to their needs. Further additions made to the system will involve a web service API as well as a domain-specific scripting language that allows users to programmatically interact with the system, to a secure degree. Many of the elements of the planned system will be, or are being, adapted from independent components commonly used by the linguistic research community, as is the case with, for instance, libraries for the statistical programming language R (Baayen, 2008), management of distributed file systems, or regex packages.

## 4. State of development

KorAP development proceeds in several parallel threads, involving testing the existing solutions and experimenting with new ones. These threads are linked by mutual feedback loops, so that agreement can be established with respect to, among others, the storage/retrieval architecture, the underlying data model, the XML data model that mediates between the user and the internal format, and the features that we wish to see in the emerging query language. An overview of the existing interface solutions is also under way, complemented by transforming the survey of our IDS colleagues' expectations and requirements into realistic use cases that guide the design and coding at various levels. Below, we briefly look at selected development threads.

### 4.1. Data storage and the underlying data model

As mentioned above, the data storage is the part of the backend in which most development takes place at this stage. While storage is not a monolithic module but involves several sub-tasks, we approach the implementation on the precondition that accessing and querying data have to scale flexibly. This is why we cannot depend on the existing approaches in which the entire corpus is loaded into memory or is stored in relational database systems (Ghodke & Bird, 2008). Under these circumstances, the challenge is to find ways to index texts and annotations in such a way that they are found and retrieved quickly.

Our approach is to use the open-source information retrieval engine Lucene as the baseline and to treat corpus querying as "a new kind of information retrieval that is sensitive to syntactic information" (Ghodke & Bird, 2010), although our platform is designed to generalise this point of view so that it is sensitive to all kinds of annotations. We concur that "unlike traditional IR systems, corpus retrieval systems not only have to deal with the 'horizontal' representation of textual data, but with heterogeneous data on all levels of linguistic description" (Schneider, 2012).

Performance-measuring methods are being developed in order to verify the ability of index-building software to handle that challenge, so that Lucene can be substituted by other methods if necessary (Figure 2). Apart from typical statistical properties found in natural language (Zipf, 1949), the benchmark will test indexing and querying performance under less likely conditions such as: few different words, many equally distributed words, very long or very short words, very large or very small documents. Despite the low probability of those distributions to occur in real corpora, this will allow us to identify logical and implementa-

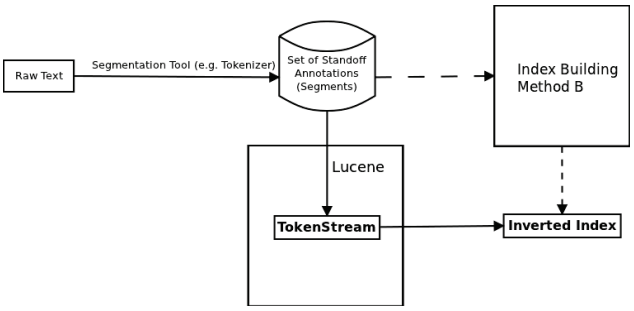


Figure 2: From raw text to an inverted index

tion flaws, and to localize potential problems that may occur when dealing with data amounts of higher magnitudes.

The main benefit of Lucene is that it is a well-established framework that provides state-of-the-art techniques for building inverted indexes from texts. While we do not need typical information retrieval methods such as stemming or stop-word filtering, we do need inverted indexes that allow for a fast look-up of tokens or other linguistic units. For that purpose, we have implemented a custom TokenStream class that extends the corresponding abstract Lucene class. This implementation reads an existing segmentation from an external file instead of performing an analysis on the input text at run-time. The resulting TokenStream integrates into the Lucene workflow so that its established indexing techniques can be applied.

The approach has been generalised so that it builds indexes for any other segmentation types as well, e. g. those based on morphemes; therefore, in what follows, we speak about segments rather than tokens. Every segmentation sequence is stored in a separate index and each of these indexes can be searched in parallel. Incoming queries are split into atomic sub-queries and distributed to the corresponding indexes. The results are joined or intersected on the basis of the segment offsets.

In a further generalisation, the implementation is also applicable to annotations that do not merely segment text but rather apply tag labels to already identified segments. In such cases, the indexed term is not a sequence of characters read from the primary text but the label, e. g. a part-of-speech tag, itself.

For range-based annotations that cannot be represented as inverted indexes in a meaningful way, for instance sentence and paragraph segmentations, we investigate the suitability of well-established techniques such as B-Trees (Comer, 1979) and Tries (Fredkin, 1960). The latter have successfully been applied to store and index n-gram based language models by Germann (2009). Presumably, the suitability of the different indexing methods largely depends on the data structure of an annotation (e. g. the vocabulary size of a tagger). In order to create an empirical base that allows us to draw conclusions about the performance of an indexing strategy, we are developing a benchmark that is able to measure and compare indexing and query performances for large scale corpora.

### 4.2. Interface and the corpus query language

Although the actual coding of the interface is a work package scheduled for a later stage of the project, the design has begun with an overview of the existing systems, and, at the same time, a survey of the expectations of the projected users of the system, followed by the construction of a set of use cases that express these expectations.

A number of existing corpus-processing systems have been analysed, with a view to comparing their strengths and weaknesses in practical application. The focus of the analysis lay on the query languages as well as the query interface design, the search functionalities, and the strategies for the presentation of the results. The goals were to identify a query language that could be re-used as the basis for the query language of the new corpus analysis platform, and to survey the successful interface features of other corpus processing systems. Among the systems that have been analysed are COSMAS II, Poliqarp (Janus & Przepiórkowski, 2007), TIGERSearch (Lezius, 2002), ANNIS (Rosenfeld, 2010), SketchEngine (Kilgarriff et al., 2004), the TXM Platform (Heiden, 2010), and DDC (Sokirko, 2003) on which both the Russian National Corpus (Plungjan, 2009) and DWDS (Geyken, 2007) are based.

While a lot of interesting and innovative functionalities and application settings have been identified, providing useful guidelines for KorAP interface development, none of the query languages that we have reviewed satisfied our requirements and expectations fully. Partial results of our query language evaluation are presented in (Frick et al., 2012).

Parallel to this work, some of the current COSMAS II users were surveyed in order to identify the needs the linguistic community may have for the new corpus processing platform. Interviewing our colleagues, grammarians and lexicographers, about their tasks and the functionalities that they miss in COSMAS II or are not comfortable with, provided us with a list of features that should or should not be implemented by KorAP.

The list of the features that the surveyed users expect or desire has been restated as collections of UML Use Case and Activity diagrams that represent many of the problems that linguists encounter in their work, and methods that are used to handle them. Another result of the survey is over 300 abstract queries expressed in natural language that have been translated into actual query languages of the corpus-analysis tools that we have tested against our test data suite, in order to determine the engine and the query language that is able to satisfy most of our goals, and at the same time, to inform the development of the new ISO TC37/SC4 Work Item “Corpus Query Lingua Franca” (Bański & Witt, 2011).

### 4.3. Canonical XML data model and test suite

While the currently projected storage format is not human-readable (see Section 4.1), the user is going to submit and retrieve annotations from KorAP via an XML layer that, on the one hand, constitutes an abstraction of the internal format, and on the other, provides a simple mapping into the XML standards commonly used for the description of language resources, viz. XCES (Ide et al., 2000), Tiger2 (Romary et al., 2011), PAULA (Chiarcos et al., 2008), the

```
<span id="s_11" from="70" to="73">
  <fs type="lex"
      xmlns="http://www.tei-c.org/ns/1.0">
    <f name="lex">
      <fs>
        <f name="lemma">die</f>
        <f name="certainty">0.930981</f>
        <f name="ctag">ART</f>
      </fs>
    </f>
  </fs>
</span>
```

Listing 1: Fragment of morphological annotation (TT)

```
<span id="s2_n0" from="70" to="73">
  <rel label="DETERM">
    <span from="74" to="85"/>
  </rel>
</span>
<span id="s2_n2" from="74" to="85">
  <rel label="SUBJ">
    <span from="117" to="130"/>
  </rel>
</span>
 ....
```

Listing 2: Fragment of dependency annotation (XIP)

ISO TC37/SC4 family (Ide & Romary, 2007) or TEI-based formats.

As an abstraction of the underlying format, the XML layer is by default span-based, though a fallback to ID-based pointing is also possible. Listing 1 illustrates a fragment of the annotation of the article das in a sentence from the test data suite derived from German Wikipedia sub-corpus (WPD) of DeReKo-2010-I (IDS, 2010). It combines information about the extent of the target span of primary text (stored in a separate document as a single sequence of Unicode characters, with no annotations whatsoever, cf. Bański and Schnober, 2012) with the ISO/TEI Feature Structure Representation (Lee et al., 2004), derived from the TreeTagger (Schmid, 1994).[3]

Listing 2 presents the output of a dependency parser (Xerox Incremental Parser), which encodes the dependency relation between the determiner 'das' and the noun 'Unternehmen', and between the noun and the main verb 'spezializiert'.[4] This is visualised in Figure 3 – a (slightly edited) fragment of the output of the ANNIS database, to which the XML format shown here was converted.

The XML test data suite is available from the project page. It contains TreeTagger annotations that we are allowed to

[3] Note, incidentally, that the lemma of das is, with a large degree of certainty, declared to be die – which only strengthens our point of the necessity to store tool output separately from the text, both for the sake of keeping the text theory-free, and for the purpose of confronting various tools in order to help improve them.

[4] XIP (`http://open.xerox.com/Services/XIPParser`) was employed experimentally within the DeReKo project; the annotations derived with it may not, for licensing reasons, be redistributed in the open Wikipedia data suite release.

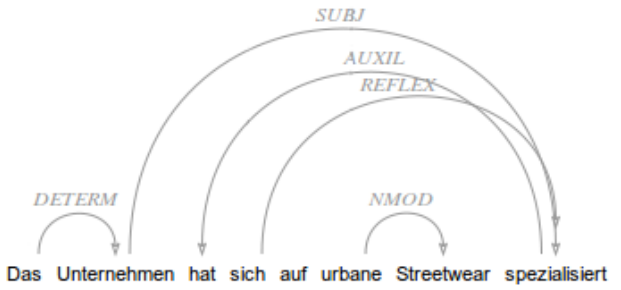


Figure 3: ANNIS visualisation of the structure containing Listing 2 (edited for readability)

publish under the same license as the primary data, with Connexor MPT[5] layers of annotation potentially shipped separately, depending on the licensing restrictions. It also accompanied by a GPL-ed validating tool. The next version of the suite is going to contain open-content dependency annotations and provide the possibility to include data-category alignment information (cf. Kemps-Snijders et al., 2008).

## 5. Summary and outlook

The present contribution has introduced the KorAP project, its aims, assumptions, and the challenges that we address or expect to face. The project is now in an early coding phase, with a lot of experimentation still on schedule. Above, we have presented the general architecture of the system and the implementation of its basic elements.

Following the principles of empiricism and open science, we are going to release as many elements of the projected system as possible. The first deliverable is version 0.2 of the test data suite, derived from German Wikipedia and equipped with TreeTagger annotations. We are going to gradually increase the number of grammatical descriptions by using other free tools.

Current information about the project will be maintained at the following URL: `http://korap.ids-mannheim.de/`

## 6. Acknowledgements

The project described here is funded within the Senate Committee Competition (SAW) programme of the Leibniz Association. We would like to thank Cyril Belica for advice and consulting, Helge Krause, and the remaining project team members Franck Bodmer and Peter Harders. We would also like to thank the anonymous reviewers of the initial short version of this paper for their valuable comments and suggestions.

## 7. References

Wadi, D.al (1994). *COSMAS - Ein Computersystem für den Zugriff auf Textkorpora*. Institut für Deutsche Sprache.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge.

[5] `http://www.connexor.com/nlplib/?q=mpt`

Bański, P. & Schnober, C. (2012). *KorAP-WPD test data set, ver. 0.2 – README*. Institut für Deutsche Sprache, March.

Bański, P. & Witt, A. (2011). Do linguists need a corpus query lingua franca? Presentation given at the ISO TC37 meeting in Seoul, South Korea, on 13 June 2011.

Belica, C., Kupietz, M., Lüngen, H. & Witt, A. (2011). The morphosyntactic annotation of DeReKo: Interpretation, opportunities and pitfalls. In Konopka, M., Kubczak, J., Mair, C., Šticha, F. & Wassner, U. (Eds.), *Selected contributions from the conference Grammar and Corpora 2009*, pages 451–471, Tübingen. Gunter Narr Verlag.

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

Broeder, D., Declerck, T., Kemps-Snijders, M., Keibel, H., Kupietz, M., Lemnitzer, L., Witt, A. & Wittenburg, P. (2007). Citation of electronic resources: Proposal for a new work item in ISO TC37/SC4. ISO TC37/SC4-Document N366. `http://www.tc37sc4.org/new_doc/ISO_TC37_SC4_N366_NP_CitER_Annex.pdf`.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J. & Stede, M. (2008). A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*, 49(2):271–293.

Comer, D. (1979). Ubiquitous B-tree. *ACM Computing Surveys (CSUR)*, 11(2):121–137.

Fiehler, R., Wagener, P. & Schröder, P. (2007). Analyse und Dokumentation gesprochener Sprache am IDS. In Kämper, H. & Eichinger, L. M. (Eds.), *Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache*, volume 40 of *Studien zur deutschen Sprache*, pages 331–365. Narr, Tübingen.

Fredkin, E. (1960). Trie memory. *Communications of the ACM*, 3(9):490–499.

Frick, E., Schnober, C. & Bański, P. (2012). Evaluating query languages for a corpus processing system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Germann, U., Joanis, E., Larkin, S. & others, (2009). Tightly packed tries: How to fit large models into memory, and make them load fast, too. In *Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009)*.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the twentieth century. In Fellbaum, C. (Ed.), *Idioms and collocations: Corpus-based linguistic and lexicographic studies*, page 23–40. Continuum, London.

Ghodke, S. & Bird, S. (2008). Querying linguistic annotations. In *Proc. 13th Australasian Document Computing Symposium*, pages 69–72.

Ghodke, S. & Bird, S. (2010). Fast query for large treebanks. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 267–275. Association for Computational Linguistics.

Gray, J. (2003). Distributed Computing Economics. Technical Report MSR-TR-2003-24, Microsoft Research.

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro, R., Ishikawa, K., Umemoto, H., Yoshimoto, K. & Harada, Y. (Eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, page 389–398, Sendai, Japan. Institute for Digital Enhancement of Cognitive Development, Waseda University. `http://halshs.archives-ouvertes.fr/halshs-00549764`.

Ide, N. & Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjær, L., Hemsen, H. & Minker, W. (Eds.), *Evaluation of Text and Speech Systems*, pages 263–284. Kluwer Academic Publishers.

Ide, N., Bonhomme, P. & Romary, L. (2000). An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 825–830.

IDS, (2010). Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010–I (released 2010-03-02). Mannheim: Institut für Deutsche Sprache. `http://www.ids-mannheim.de/dereko`.

Janus, D. & Przepiórkowski, A. (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*, pages 85–88. Association for Computational Linguistics.

Keibel, H., Kupietz, M. & Belica, C. (2008). Approaching grammar: Inferring operational constituents of language use from large corpora. In Štícha and Fried (Štícha and Fried, 2008), page 235–242.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P. & Wright, S. (2008). ISOcat: Corralling data categories in the wild. In *Proceedings of the International Conference on Language Resources and Evaluation, Marrakech, Morocco, May*.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The sketch engine. *Information Technology*, 105:116.

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, 33(1):147–151.

Kupietz, M. & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In Minegishi, M. & Kawaguchi, Y. (Eds.), *Working Papers in Corpus-based Linguistics and Language Education, No. 3*, page 53–59. Tokyo University of Foreign Studies (TUFS), Tokyo. `http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf` (12.06.2009).

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M. & Tapias, D. (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, page 1848–1854, Valletta, Malta, May. European Language Resources Association (ELRA). `http://www.lrec-conf.`

org/proceedings/lrec2010/pdf/414_Paper.pdf (25.5.2010).

Lee, K., Burnard, L., Romary, L., Bunt, H. & Clergerie, E. (2004). Towards an international standard on feature structure representation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 373–376.

Lezius, W. (2002). TIGERSearch – Ein Suchwerkzeug für Baumbanken. In *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, volume 6, pages 107–114.

Lüdeling, A. & Kytö, M. (2008). *Corpus linguistics: an international handbook*. Handbücher zur Sprach- und Kommunikationswissenschaft. de Gruyter, Berlin.

Müller, S. (2007). Qualitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus. In Zifonun, G. & Kallmeyer, W. (Eds.), *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*, Institut für Deutsche Sprache Jahrbuch 2006, pages 70–90. de Gruyter, Berlin.

Neuroth, H., Lohmeier, F. & Smith, K. M. (2011). Textgrid – virtual research environment for the humanities. *The International Journal of Digital Curation*, 6(2). (Proceedings of the 6th International Digital Curation Conference, Chicago, USA, Dec 2010).

Pedersen, T. (2008). Empiricism Is Not a Matter of Faith. *Computational Linguistics*, 34(3):465–470. http://www.d.umn.edu/~tpederse/Pubs/pedersen-last-word-2008.pdf.

Plungjan, A. M. (2009). *Национальный корпус русского языка*. Нестор-История.

Romary, L., Zeldes, A. & Zipser, F. (2011). [Tiger2/]-Serialising the ISO SynAF Syntactic Object Model. *Arxiv preprint arXiv:1108.0631*.

Rosenfeld, V. (2010). An implementation of the annis 2 query language. Technical report, Humboldt-Universität zu Berlin.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, page 44–49, Manchester, UK.

Schneider, R. (2012). Evaluating DBMS-based access strategies to very large multi-layer corpora. In *Proceedings of the LREC 2012 Workshop: Challenges in the management of large corpora*. European Language Resources Association (ELRA).

Sinclair, J. (1994). Trust the text. In Coulthard, M. (Ed.), *Advances in written text analysis*, pages 12–25. Routledge, London.

Sokirko, A. (2003). *A technical overview of DWDS/Dialing Concordance*.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Benjamins, Amsterdam, Philadelphia.

Štícha, F. & Fried, M. (Eds.) (2008). *Grammar & Corpora 2007, Selected contributions from the conference Grammar and Corpora*. Academia, Prag.

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M. & Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S. & Tapias, D. (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Zipf, G. (1949). Human behavior and the principle of least effort: An introduction to human ecology. *AddisonWesley Press*.