Recent Developments in DeReKo

Marc Kupietz, Harald Lüngen
Institute for the German Language (IDS)
R5 6–13, 68161 Mannheim, Germany
{kupietz|luengen}@ids-mannheim.de

Abstract

This paper gives an overview of recent developments in the German Reference Corpus DeReKo in terms of growth, maximising relevant corpus strata, metadata, legal issues, and its current and future research interface. Due to the recent acquisition of new licenses, DeReKo has grown by a factor of four in the first half of 2014, mostly in the area of newspaper text, and presently contains over 24 billion word tokens. Other strata, like fictional texts, web corpora, in particular CMC texts, and spoken but conceptually written texts have also increased significantly. We report on the newly acquired corpora that led to the major increase, on the principles and strategies behind our corpus acquisition activities, and on our solutions for the emerging legal, organisational, and technical challenges.

Keywords: German reference corpus, very large corpora, legal issues

1. Introduction

We report on recent developments in DeReKo, the *Archive of General Reference Corpora of Contemporary Written German* hosted at the Institut für Deutsche Sprache in Mannheim (IDS). DeReKo is designed as a very large general-purpose corpus archive. Its primary purpose is to serve as an empirical basis for linguistic research on contemporary written German, i.e. it is not designed specifically for lexicography, but for all fields of language research where quantitative analyses requiring large corpora are conducted.

DeReKo is probably the largest linguistically motivated archive of German texts. It contains fictional texts, newspaper texts, specialised texts, scripted speech, internet-based communication, and many other text types. DeReKo is continually being expanded. Only complete texts are included, originating from around 1956 or later. DeReKo is annotated on multiple linguistic levels, e.g. with POS tagging and syntactic dependency structures.

Unlike other reference corpora, DeReKo is not designed to be balanced in any way, because what kind of balance is appropriate always depends on the research question and the language domain under scrutiny. We think that researchers ideally should be able to answer questions like *Which language domain do I want to examine? Is 20% or 30% or 50% fictional texts to be considered as balanced? Which language strata are relevant?* themselves and should have means at their disposal to compile a suitable corpus accordingly. Hence, DeReKo is designed to serve as a *primordial sample* of language use, from which users can draw stratified *virtual corpora* that are representative or balanced w.r.t. their research question (cf. Kupietz and Keibel, 2009; Kupietz et al., 2010). The primordial sample design ensures an optimal usability of the corpus data for the maximum number of potentially relevant research questions. It also allows for an optimisation of the cost-benefit ratio, e.g. data offered to us for free need not be declined.

The present paper reports on newly acquired corpora that led to a recent major increase of DeReKo, on the principles and strategies behind our corpus acquisition activities, and on our solutions for the emerging legal, organisational, and technical challenges. While our general aim is to inform the scientific community about current developments and results, we also intend this paper as a contribution to the ongoing specialised exchange among the national corpora and very large reference corpora initiatives (e.g. Geyken, 2007; Przepiórkowski et al., 2010; Ransmayr et al., forthcoming).

2. Legal issues

The IDS does not own any of the texts contained in DeReKo. In order to be able to use them and make them available to the scientific community, we had to conclude license contracts with the respective copyright holders. The rights of use formulated in these licenses have some limitations, most importantly that a.) only academic use is allowed whereas commercial use is explicitly forbidden, b.) access is a only allowed via specialised query software, which technically prevents the reconstruction, let alone download of complete texts, and c.) only authenticated users may be granted access. Note that without such compromises, we would not be able to compile a corpus like DeReKo, as less restrictive license terms are extremely expensive. Currently, DeReKo contains texts from more than 200 donors (publishing houses or individual authors).

By the two major types of agreements, the corpora are made available for IDS employees and guests only, or made available world-wide. To be able to specify the complex rights situations, we propose one new *main category* and four new *additional modifiers* as an extension of the CLARIN system originally proposed by Oksanen et al. (2010). As a new main category besides PUB, ACA, and RES, we propose *QAO* (query and analysis only):

QAO – not downloadable, the end user has the right to query and analyse the resource for academic purposes via specialised software provided by the copyright curator

As four new *additional modifiers,* which can in principle be combined with any main category, we propose:

LOC:*loc* – the resource may not be copied outside the servers of the copyright curator, e.g. LOC:ids

QAO-NC	academic, non-commercial, query-and-analysis only (i.e. accessible only via COSMAS-II) use
QAO-NC-LOC:ids	academic, non-commercial, query-and-analysis-only (i.e. accessible only via COSMAS-II) use, only at the site of the IDS
QAO-NC-LOC:ids-NU:1	academic, non-commercial, query-and-analysis-only (i.e. accessible only via COSMAS-II) use, only at the site of the IDS, only by one user at a time
ACA-NC	academic, non-commercial use, no re-distribution by the end user
ACA-NC-LC	academic, non-commercial use, no re-distribution by the end user, license contract ŵith copyright holder required
CC-BY-SA	other e.g. Wikipedia

Table 1: Combinations of CLARIN and newly proposed laundry symbols used in DeReKo

LC – a signed license contract with the copyright holder is required

NU:*n* – the resource can only be made available to *n* end users at a time

TER – a territorial restriction, e.g. TER:Germany for use only in Germany

According to this proposal, each DeReKo subcorpus is now specified with one of the values listed in Table 1 in the <availability> element of its TEI header.

We are convinced that these or similar additions to the CLARIN classification system would raise its coverage, granularity and usefulness significantly, as, because of the affected third parties' rights, almost all contemporary corpora are not available for download, but only accessible, searchable and analysable through specialised software (cf. Hinrichs and Beck, 2011, p. 47ff). Apart from being a legal and economical reality and necessity in most European countries, making corpora available in such a way makes also sense from a computer scientific point of view, ideally putting Jim Gray's (2003) famous postulate *put the computation near the data* and one of the founding motivations for CLARIN, namely *getting away from the download-first paradigm,* into practice. We are of course aware of the implications on the responsibilities of corpus curators and providers to make the data, even though it cannot be downloaded, as useful for researchers, as legally and practically possible (see section 6.).

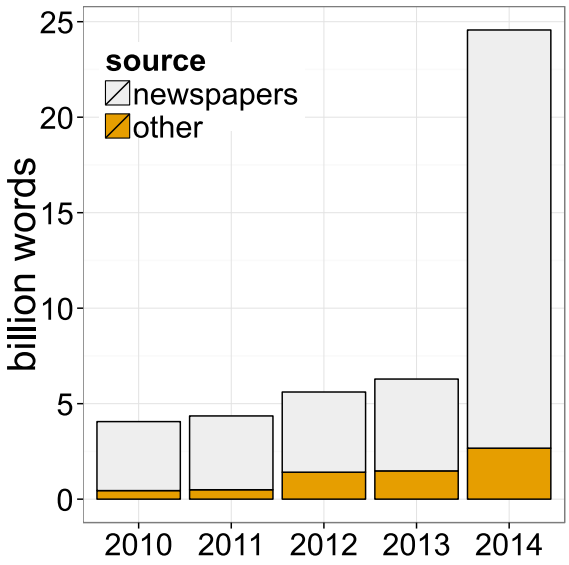


Figure 1: DeReKo growth since 2010

3. Growth

DeReKo is continually expanded and presently one of the biggest corpus archives of texts in (contemporary) German. It contained over four billion running words in 2010, and as a result of recent campaigns and major acquisition deals, DeReKo has grown to over six billion running words in 2013 and has further grown by a factor of four in the first half of 2014 to more than 24 billion words (see Figure 1). DeReKo's present growth rate (considering the regularly incoming text data) is 1.7 billion words per year. Note that size is not an end in itself. Language is known to contain a large number of rare events – not only lexical events, but also events defined by a combination of conditions. The bigger a corpus is (while other parameters remain the same), the more reliable conclusions can be drawn from it about rarer and more diversified phenomena.

3.1. Recent campaigns and acquisitions

With respect to DeReKo's design as a primordial sample, activities to expand DeReKo further are guided by two principles: maximisation of size and maximisation of dispersion with respect to potentially relevant strata. In practice, further criteria play a role, such as demand from IDS-internal and -external projects, supply by right holders or projects in which corpora are built, and in particular, licensing costs and the costs of preparing the raw data for integration in DeReKo.

Unlike in the age of the BNC (Aston and Burnard, 1998), major discrepancies among different text genres and sources w.r.t. these criteria have arisen especially in the last few years. Newspaper publishing companies have focussed early on digital editions and converted their production chains to single source publishing, starting with advanced authoring systems and ending with print editions, various e-paper formats, and export formats for digital archiving, from which DeReKo can frequently benefit. German book publishers, however, have joined this trend only very recently and slowly.

Hence, a major part of the continual growth of DeReKo has been due to the regular supply of the latest digital editions of 15 German language newspapers according to existing agreements with the publishers, and the genre of newspaper text has traditionally been a prevailing stratum in DeReKo. Due to the primordial sample design, relative differences of the sizes of strata are not relevant in the case of DeReKo, as virtual corpora with the desired proportions can always be drawn. What matters, however, in this context are absolute sizes – especially in the case of sparsely populated strata.

3.1.1. Fiction

One of the weaker strata in DeReKo has been the stratum of fictional texts, and to develop it further, our 2011 acquisition campaign focused on fiction. 69 publishers of fiction in German-speaking countries were contacted. Eleven of them answered that they were willing to support DeReKo, and further negotiations resulted in eight new license agreements, in which the respective publishers would grant free licenses for the use of works of fiction in DeReKo. Six of them actually sent text data in the end. While most publishers grant a selection of 10 to 20 books, two of the six granted all titles from their respective backlists. Consequently, the subcorpora of fictional texts in DeReKo have increased by 50% since 2010 and presently comprise 17.6 million word tokens.

The books were delivered in formats as diversified as EPUB, InDesign, and PDF. As the markup of these formats is mostly layout-oriented and varies much even among the texts from the same publishing house, a considerable amount of labour had to be invested into developing new converter modules and adapting existing conversion pipelines to convert the data into the IDS text model. Due to these circumstances, the expenses for the acquisition and curation of one word of fictional text are presently about 25,000 times higher than the expenses for one word of newspaper text.

3.1.2. Web-based corpora

Several recent big data corpus initiatives have introduced new methods to derive clean text from web pages and have improved POS-tagging to deal with the peculiarities of web language and web documents (Jakubíček et al., 2013; Schäfer and Bildhauer, 2012; Baroni et al., 2009). However, due to DeReKo's quality standards to include only licensed material and to provide sufficient metadata for enabling users to derive virtual subcorpora, we cannot readily use the web as a source. It is hard to impossible to obtain or derive useful metadata for web documents, not even basic ones such as authorship, let alone the time of composition, the author's native language or geographical affiliation (cf. also Jakubíček et al., 2013). For many web documents, it is hard to determine whether their text content has been composed by a human at all. Moreover, German and European copyright laws lack the notions of *fair use* and *implied license*, hence an explicit license would have to be obtained from every single author of a candidate webpage or forum entry, which seems infeasible (see also Guevara, 2010). For these reasons, our strategy to cover specific web genres in DeReKo is to focus on web archives and collections that have been published under sufficiently liberal licenses, or where there is a central authority, such as a major forum or blog host, with whom we can negotiate a comprehensive license. Under these preconditions, we have adapted the complete archive of German Wikipedia articles and discussion pages for DeReKo in 2011 (Bubenhofer et al., 2011), and again in 2013 (Margaretha and Lüngen, in preparation)[1]. The 2011 Wikipedia conversion comprises 830 million word tokens and formed the bulk of DeReKo's increase in 2012. The 2013 conversion comprises more than 1 billion word tokens.

Wikipedia discussions are an instance of *computer-mediated communication* (CMC). CMC represents written, but conceptually spoken language and is an important manifestation of contemporary language (Lemnitzer et al., 2012). Within a new project on orthographic usage, started in 2013 in cooperation with the Council for German Orthography, we currently seek to acquire further CMC data from web sources such as forums, newsgroups, and blogs under the premises sketched above.

3.1.3. Conceptually written language

In 2012, we also extended the stratum of medially spoken, but conceptually written language. We acquired the German Political Speeches Corpus compiled by Barbaresi (2012)[2], and the parliamentary debates corpus PolMine[3], comprising the complete protocols from the sessions of the 18 German national and state-level parliaments since 2000. The conversion tools used for the PolMine corpus were in parts developed in a co-operation between the IDS and the PolMine group of the University of Duisburg-Essen and are employed for regular updates of this corpus with the newest protocols. We intend to also curate the debate protocols from the parliaments in Austria, and German-speaking Switzerland and Belgium, as well as the protocols from before the year 2000. The political speeches and PolMine corpora within DeReKo presently comprise 316 million running words.

3.1.4. News database archive

In 2013, we struck a deal with a commercial German-language news database provider, according to which DeReKo obtained licenses for about 102 million documents (69 GB zipped XML) of press text, specialised journals and e-books. The size of the archive posed major challenges for the technical infrastructure of DeReKo and for the corpus extension project.

All documents came marked up according to the provider's own XML format, which also included metadata specifying for each document the source, issue, date, page number, title, and (partly) one or more domain categories and one or more keywords (mostly names of people or locations). The basic text structure markup exhibited some variation w.r.t. to the marking of paragraphs, headings, and subheadings. For the latest DeReKo release DeReKo-2014-I, we selected the press data part which contained consecutive editions of 98 national and regional newspapers and magazines, mostly

[1] also available for download, see http://www.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html

[2] see http://perso.ens-lyon.fr/adrien.barbaresi/corpora

[3] http://polmine.sowi.uni-due.de/

starting between 2000 and 2003.

Preparation As a pre-processing step, we applied some standard filters, i.e. discarding texts that had the same ID or the same md5 checksum as a previously seen text, texts containing less than ten words, texts with more than 10% digits, and finally texts with less than 15% function words. (In the future, we will also apply our proper duplicate detection and encode the duplicate and near duplicate relations in the corpus files as described in Kupietz (2005) and Klosa et al. (2012)).

We designed an XSLT stylesheet to convert the XML sources into the I5 markup (see Section 5.), applying heuristics to extract further metadata such as the author of an article, to deal with the variation in text structure and to mark up regions of text in more detail than the original, e.g. additionally annotating openers and bylines. We also applied our topic classification (see Weiß, 2005; Klosa et al., 2012) and sentence splitting routines and added the respective results to the I5 encoding.

For quality assurance (in addition to the above described filtering), we manually inspected a sample of documents chosen from the set of the 12 major papers. As a result, we further adjusted the functions for identifying author and subtitle towards a better precision.

Eventually, the resulting corpus ready for inclusion in DeReKo comprised more than 70 million documents (i.e. news articles), containing more than 16 billion word tokens and using up more than 380 GB of disk space. The entire processing (i.e. filtering, XSL transformation, XML validation, and word count) ran in 13 parallel threads on 48 cores at 50% CPU load and took 14 hours altogether.

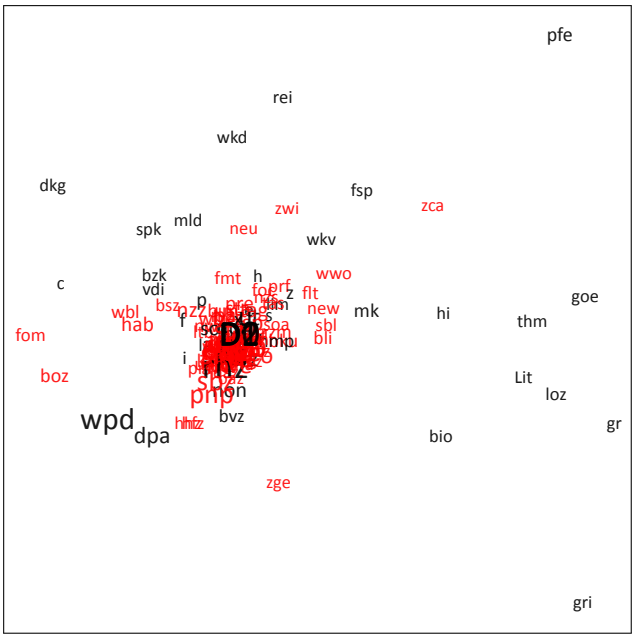
Usually, we provide DeReKo data with linguistic annotations on three syntactic layers as standoff annotation (see Belica et al., 2011), but DeReKo with the annotations has always been bigger by a factor of 40 i.e. would amount to 15.2 TB for DeReKo including the new corpora. Since presently we do not have that much storage space available, we have provided only a part of the release DeReKo-2014-I with linguistic annotations for the time being. For the future, we have ordered 24 TB additional disk space.

4. Internal variance in the new DeReKo release

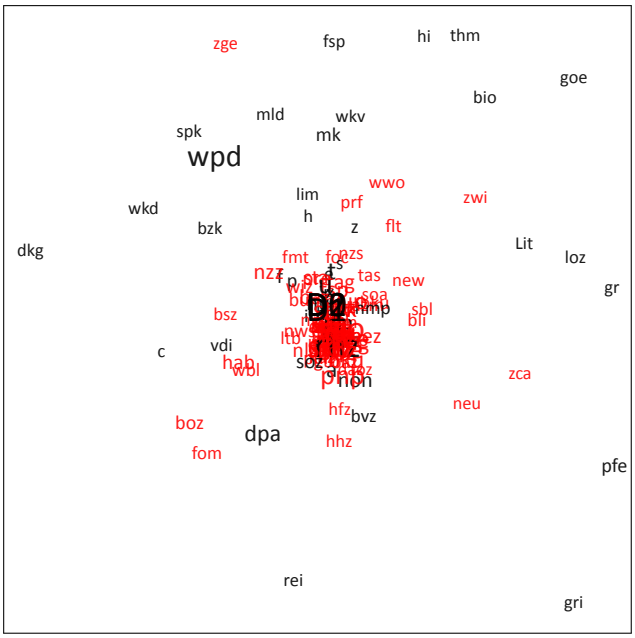
The release DeReKo-2014-I contains over 24 billion word tokens altogether (in contrast to the previous release DeReKo-2013-II which contained 6.6 billion tokens). Since we knew that the present IDS' corpus search, management and analysis system COSMAS II cannot handle an archive of this size with reasonable indexing and query response times, we thought of ways to prioritise the new data to allow for choosing the "most important" subcorpora for inclusion in COSMAS, so that they can be used even before the launch of our new and more powerful corpus platform KorAP (see Section 6. and Bański et al., 2014, in this volume). Since in the primordial sample design it is always desirable to increase the internal variance of the primordial sample, we tested all subcorpora for similarity/distance to each other and to the old DeReKo (DeReKo-2013-II) as a whole, using Kilgariff's (2001) word frequency list-based distance measures for comparing corpora. The intention was to preferably include the most dissimilar of the new subcorpora in COSMAS as they would increase variance most, and among them those with the best coverage of strata that are usually under-represented w.r.t. to typical virtual corpus definitions, such as fiction and history (as opposed to finance and local announcements). For each subcorpus to be compared, we derived random samples of 1 million tokens, and to be on the safe side, we derived three such samples from DeReKo-2013-II as a whole. Then for each pair of samples to be compared, we took the lists of the 500 most frequent words in the union of both samples according to Kilgariff's (2001) method and calculated their similarity/distance using Kilgariff's χ^2-based measure. We transformed the resulting distance matrix into a two-dimensional map using non-metric multidimensional scaling (NMDS), see Figure 2a. The three DeReKo-2013-II samples (D0,D1,D2) got located next to each other in the centroid of the map, reflecting that they are the most average samples, and showing that sampling, distance measure, and projection were sufficiently robust for our purposes. Each new subcorpus from the newly acquired corpora is shown in red, and each old subcorpus is shown in black. The most distant among the old corpora are corpora containing transcribed spoken language (pfe), fairy tales (gri), fiction (thm, goe, loz, …), political speeches (rei), and Wikipedia (wpd)[4]. The major groups of distant "outliers" from the newly acquired corpora are made up by domain-specific magazines (zca, zge, zwi, flt, …), business newspapers and mags (fom, boz), and local papers (hhz, hfz, pnp). To check for possible biases e.g. caused by region-related names which might be dominating in certain samples and to focus away from domain aspects in the direction of register and text type aspects, we made the same calculations based on frequency lists containing adjectives only (Figure 2b), and conjunctions only (Figure 2c) according to our TreeTagger-POS-annotations (Schmid, 1994; Belica et al., 2011). With respect to our purpose of getting an idea of which additions to the part of DeReKo that can be made available with COSMAS II would most effectively increase the dispersion in the direction of the desired strata, the results were, however, quite equivalent, all having zge, zca, zwi, neu, flt, wwo among the first candidates. In order to verify our interpretations of the axes and regions of the previous NMDS-projections we eventually also computed the pairwise distances of the topic-classification distributions in the DeReKo-samples and the samples of the newly acquired sources (Figure 2d).

Another strategy to increase the variance in the archives made available in COSMAS would be to preferably include newspaper corpora from locations that increase the dispersion in the geographical distribution of sources most. As can be seen in Figure 3, the newly acquired corpora will help close gaps in the coverage of DeReKo of the German-speaking areas, especially of Austria, Switzerland, Luxembourg, and the west and east of Germany, still leaving a gap in the north.

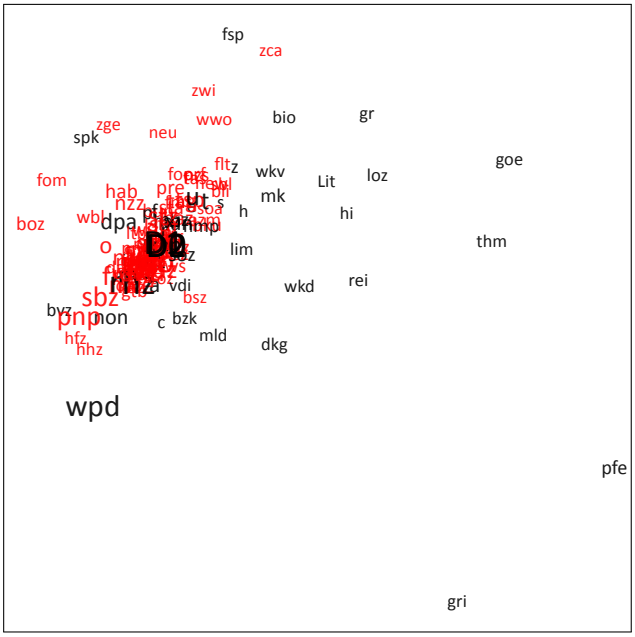
[4] See http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html for a comprehensive explanation of all abbreviations.



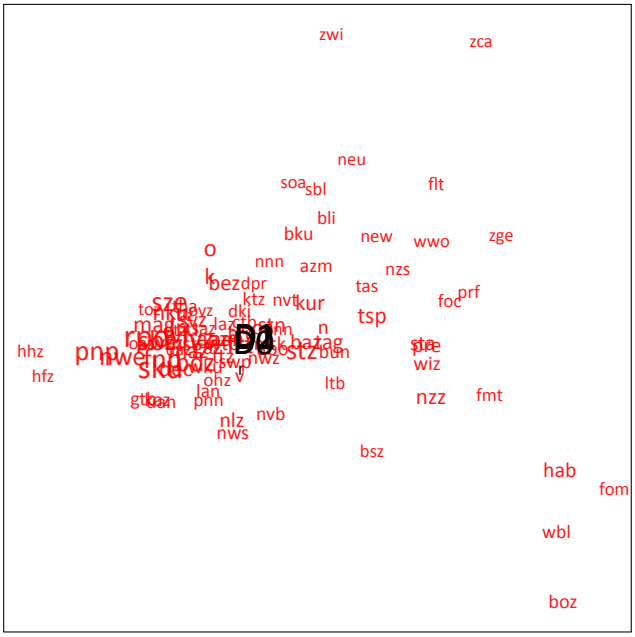
(a) based on all tokens



(b) based on adjectives only



(c) based on conjunctions only



(d) based on topic domain classifications

Figure 2: NMDS projections of χ^2-distance matrices of old (black) DeReKo-sources and new acquisitions (red) based on different frequency lists.

5. Text model, metadata and annotations

The hierarchical corpus and text structure of DeReKo is defined in the *IDS text model*. It had been realised as an IDS-specific adaptation of the Corpus Encoding Standard XCES (Ide et al., 2000) since around 2000, called IDS-XCES, which has also been the internal representation format in the IDS corpus research interface COSMAS (cf. Section 6.). (X)CES itself had been based on the TEI P3 model, restricting it to the application to linguistic corpora. With the advent of TEI P5 and the new ODD mechanism for TEI customisations, it became possible to specify formally how the IDS text model corresponds to the TEI and in exactly what points it deviates. Thus in 2012, we introduced and migrated to a new document grammar called *I5*, specified as an ODD document defining the IDS text model as a TEI P5 customisation (Lüngen and Sperberg-McQueen, 2012).

On the occasion of the 2013 Wikipedia conversion, we additionally introduced elements for the suitable representation of the thread and postings structure of contributions to CMC documents, according to the TEI proposal by Beißwenger et al. (2012).

There are no TEI P5 elements and attributes specially designed for the representation of debate protocols, so that we

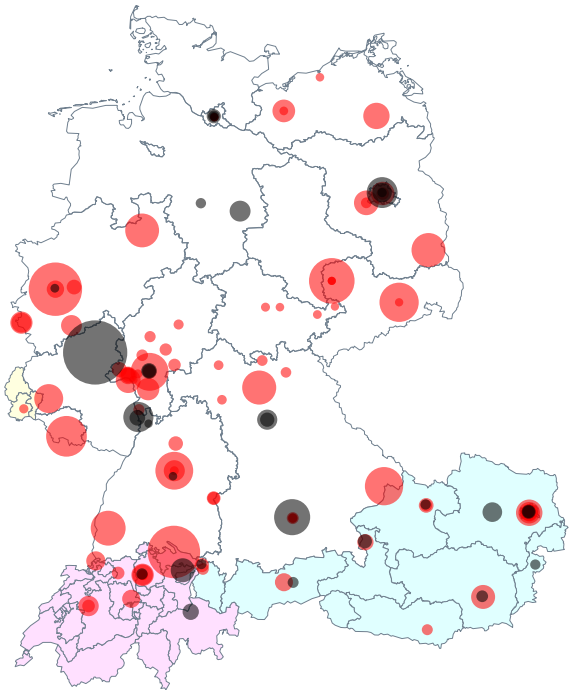


Figure 3: Geographical distribution and size of newspaper sources. Old sources shown in black, newly acquired sources shown in red.

decided to annotate the parliamentary debate protocols described under 3.1.3. according to the markup from the TEI P5 performance text module that is also available in I5, e.g. using <sp> and <speaker> for the oral contribution and the name or ID of a speaker, respectively, and <stage> for mentions of extra-linguistic events such as acclamation, but also for interjections. Due to the performative character of a debate, the markup could be used in a straightforward way. Within the metadata section (idsHeader), we now specifiy the relevant combination of "laundry tags" to describe the type of license (see Section 2.) in the <availability> element.

6. Access

Since 1993 DeReKo can be accessed free of charge (for academic use) via the Corpus Search, Management and Analysis System COSMAS (al Wadi, 1994; Bodmer, 2005; Bodmer Mory, 2014; IDS, 1991 2014). It is currently used by more than 32,000 registered users from all over the world and has been actively developed in the past years. Recent additions include e.g.,

- the ability to handle multiple morphosyntactic annotation layers
- a GUI assistant to help constructing queries based on these
- improved possibilities for the construction of virtual corpora
- query language extensions concerning distance operators and regular expressions over part-of-speech sequences
- result views with optional random order and optional break down by metadata categories like topic and text type
- a web service API, that allows access, e.g., from CLARIN and TextGrid

As the design and important parts of the code base of COSMAS II however date back to the early nineties and new developments as well as coping with the enormous growth of DeReKo becomes more and more expensive, in 2011 we have started to develop the new corpus platform KorAP from scratch. Being horizontally scalable, KorAP will support an in principle unlimited number of tokens and annotation layers (for details, see Bański et al., 2012, 2013, 2014). Its public beta release is scheduled for summer 2014.

7. Conclusion and prospects

We have given an overview of recent developments in the German Reference Corpus DeReKo in terms of growth, maximising relevant corpus strata, metadata, legal issues, and its current and future research interface. Due to the recent acquisition of new licenses, DeReKo has grown by a factor of four in the first half of 2014, mostly in the area of newspaper text, and presently contains over 24 billion word tokens. Other strata, like fictional texts, web corpora (in particular CMC), and spoken but conceptually written texts have also increased significantly, though their share in DeReKo is still relatively low due to the present conditions of their acquisition and curation. Through the latest acquisitions, the supply of newspaper archives seems to have reached a ceiling, and we will be able to allocate more resources to curating texts from other genres in the future.

We described recent additions to I5, the TEI customisation of the IDS text model. In the future, there will be a need to harmonise I5 with the text models of other institutions which are also using variants of the TEI P5 and with which we collaborate within the EU CLARIN framework (CLARIN-D AP-5, 2012). For this purpose, we are currently developing *Igel,* a web-based application for examining and comparing a collection of document grammars and deriving an overview of their differences and similarities (Sperberg-McQueen et al., 2013).

In connection to CLARIN, we will also rebase our metadata export for OAI-PMH (OAI-PMH, 2008) to our centre for the long-term-preservation of German linguistic research data which is currently being established at the IDS (Fankhauser et al., 2013). The metadata will then be exported as CMDI records (Broeder et al., 2011) for all three levels of granularity within DeReKo (corpus, document, text).

Furthermore, more linguistic annotation layers, such as syntactic dependency and constituency analyses, will be provided when KoRAP is introduced as the new research interface for DeReKo. As already envisaged in Kupietz et al. (2010), we still plan to make available some pre-defined virtual corpora with frequently requested properties and proportions in cooperation with projects of the lexis and grammar departments of the IDS. We would also be happy to cooperate with external projects partners who are interested this task.

8. References

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

Aston, G. and Burnard, L. (1998). *The BNC Handbook*. Edinburgh University Press.

Bański, P., Diewald, N., Hanl, M., Kupietz, M., and Witt, A. (2014). Access Control by Query Rewriting: the Case of KorAP. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik. European Language Resources Association (ELRA). in this volume.

Bański, P., Fischer, P. M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, O., and Witt, A. (2012). The New IDS Corpus Analysis Platform: Challenges and Prospects. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul. European Language Resources Association (ELRA).

Bański, P., Frick, E., Hanl, M., Kupietz, M., Schnober, C., and Witt, A. (2013). Robust corpus architecture: a new look at virtual collections and data access. In Hardie, A. and Love, R., editors, *Corpus Linguistics 2013 Abstract Book*, pages 23–25, Lancaster. UCREL. `http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf`.

Barbaresi, A. (2012). German political speeches, corpus and visualization. 2. version. Technical report, ENS Lyon. `http://perso.ens-lyon.fr/adrien.barbaresi/corpora/technical-paper_v2.pdf`.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226. `http://dx.doi.org/10.1007/s10579-009-9081-4`.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. (2012). A TEI schema for the representation of computer-mediated communication. *Journal of the Text Encoding Initiative*, 3.

Belica, C., Kupietz, M., Lüngen, H., and Witt, A. (2011). The morphosyntactic annotation of DeReKo: Interpretation, opportunities and pitfalls. In Konopka, M., Kubczak, J., Mair, C., Šticha, F., and Wassner, U., editors, *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.

Bodmer Mory, F. (2014). Mit COSMAS II »in den Weiten der IDS-Korpora unterwegs«. In Steinle, M. and Berens, F. J., editors, *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, page 376–385. Institut für Deutsche Sprache, Mannheim.

Broeder, D., Schonefeld, O., Trippel, T., van Uytvanck, D., and Witt, A. (2011). A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage : The Markup Conference 2011*, volume 7 of *Balisage Series of Markup Technologies*.

Bubenhofer, N., Haupt, S., and Schwinn, H. (2011). A comparable Wikipedia corpus: From Wiki syntax to POS Tagged XML. In Hedeland, H., Schmidt, T., and Wörner, K., editors, *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, volume 96B of *Working Papers in Multilingualism*, pages 141–144, Hamburg. Hamburg University.

CLARIN-D AP-5 (2012). *CLARIN-D User Guide*. CLARIN. `http://de.clarin.eu/de/sprachressourcen/benutzerhandbuch.html`.

Fankhauser, P., Fiedler, N., and Witt, A. (2013). Forschungsdatenmanagement in den Geisteswissenschaften am Beispiel der germanistischen Linguistik. *Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB)*, 60(6):296–306.

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

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

Guevara, E. R. (2010). NeWac: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010. Sixth Web as Corpus Workshop*, pages 1–7. Association for Computational Linguistics.

Hinrichs, E. and Beck, K., editors (2011). *D-SPIN: Deutsche Sprachressourceninfrastruktur – Schlussbericht*. Abteilung Allgemeine Sprachwissenschaft und Computerlinguistik; Seminar für Sprachwissenschaft; Universität Tübingen.

Ide, N., Bonhomme, P., and Romary, L. (2000). XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference (LREC'00)*, Paris. European Language Resources Association (ELRA).

IDS (1991–2014). COSMAS I/II Corpus Search, Management and Analysis System. `http://www.ids-mannheim.de/cosmas2/`.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013). The TenTen corpus family. In *Abstract Book of the 7th International Corpus Linguistics Conference CL2013*, pages 125–127, Lancaster.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Klosa, A., Kupietz, M., and Lüngen, H. (2012). Zum Nutzen von Korpusauszeichnungen für die Lexikographie. *Lexicographica*, 28:71–97. Berlin/New York: de Gruyter.

Kupietz, M. (2005). Near-Duplicate Detection in the IDS Corpora of Written German. Technical Report kt-2006-01, Institut für Deutsche Sprache. `ftp://ftp.ids-mannheim.de/kt/ids-kt-2006-01.pdf`.

Kupietz, M., Belica, C., Keibel, H., and 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., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, page 1848–1854, Valletta, Malta. European Language Resources Association (ELRA). `http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf` (24.10.2013).

Kupietz, M. and Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In Minegishi, M. and Kawaguchi, Y., editors, *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).

Lüngen, H. and Sperberg-McQueen, C. M. (2012). A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, 3:1 – 18.

Lemnitzer, L., Beißwenger, M., Ermakova, M., Geyken, A., and Storrer, A. (2012). DeRiK: A German Reference Corpus of Computer-Mediated Communication. In *Conference abstracts of the Digital Humanities 2012*, pages 259–263, Hamburg. Hamburg University Press.

Margaretha, E. and Lüngen, H. (in preparation). Building linguistic corpora from wikipedia articles and discussions.

OAI-PMH (2008). The open archives initiative protocol for metadata harvesting (oai-pmh). `www.openarchives.org/OAI/openarchivesprotocol.html`.

Oksanen, V., Lindén, K., and Westerlund, H. (2010). Laundry symbols and license management: Practical consideratios for the distriction of lrs based on experiences from clarin. In *Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*.

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2010). Recent Developments in the National Corpus of Polish. In Calzolari, N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ransmayr, J., Mörth, K., and Ďurčo Matej (forthcoming). Linguistic variation in the austrian media corpus. dealing with the challenges of large amounts of data. In *Proceedings of the International Conference on Corpus Linguistics (CILC)*, University of Alicante.

Schäfer, R. and Bildhauer, F. (2012). Building large corpora from the web using a new effcient tool chain. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, page 486–493, Istanbul. ELRA.

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.

Sperberg-McQueen, C. M., Schonefeld, O., Kupietz, M., Lüngen, H., and Witt, A. (2013). Igel: Comparing document grammars using XQuery. In *Proceedings of Balisage: The Markup Conference 2013*, volume 10 of *Balisage Series on Markup Technologies*, Montreal. doi:10.4242/BalisageVol10.Schonefeld01.

Weiß, C. (2005). *Die thematische Erschließung von Sprachkorpora*, volume 1 of *OPAL - Online publizierte Arbeiten zur Linguistik*. Institut für Deutsche Sprache, Mannheim.