Maximizing the Potential of Very Large Corpora: 50 Years of Big Language Data at IDS Mannheim

Marc Kupietz, Harald Lüngen, Piotr Bański, Cyril Belica
Institute for the German Language (IDS)
R5 6–13, 68161 Mannheim, Germany
corpuslinguistics@ids-mannheim.de

Abstract

Very large corpora have been built and used at the IDS since its foundation in 1964. They have been made available on the Internet since the beginning of the 90's to currently over 30,000 researchers world-wide. The Institute provides the largest archive of written German (Deutsches Referenzkorpus, DeReKo) which has recently been extended to 24 billion words. DeReKo has been managed and analysed by engines known as COSMAS and afterwards COSMAS II, which is currently being replaced by a new, scalable analysis platform called KorAP. KorAP makes it possible to manage and analyse texts that are accompanied by multiple, potentially conflicting, grammatical and structural annotation layers, and is able to handle resources that are distributed across different, and possibly geographically distant, storage systems. The majority of texts in DeReKo are not licensed for free redistribution, hence, the COSMAS and KorAP systems offer technical solutions to facilitate research on very large corpora that are not available (and not suitable) for download. For the new KorAP system, it is also planned to provide sandboxed environments to support non-remote-API access "near the data" through which users can run their own analysis programs.[†]

Keywords: very large corpora, scalability, big data

1. History of corpora and corpus technology at the IDS

While the IDS was founded in 1964, at least from 1967, under the directors Paul Grebe and Ulrich Engel, a department called *Documentation of the German language* was in place, in which a text collection of contemporary German was compiled and recorded on punchcards (Teubert and Belica, 2014, p.300). The first electronic corpus to be released was the Mannheimer Korpus I (MK I, 1969), which comprised 2.2 million words in 293 texts of mostly fiction, including some popular fiction, and newspaper text. In 1972, a smaller, additional part called MK II with more text types was added. In the subsequent project *Grundstrukturen der deutschen Sprache*, the MK data were extensively analysed and used as the empirical basis in 17 published volumes on grammatical themes between 1971 and 1981 (Teubert and Belica, 2014, 301). Over the years, more corpora were added, amongst other things from branches of the IDS which were hosting specific projects, such as the Bonner Zeitungskorpus (Hellmann, 1984). At that time, the corpus data were maintained by the computing centre of the IDS, and linguists had to specify their queries to programmers who would then formulate them in machine-readable form. Between 1982 and 1992, the first proprietary concordancer REFER was in use at the IDS computing centre. REFER supported interactive, sentence-oriented queries in up to 17 million running words including basic grammatical categories, e.g. verb and adjective inflection. In 1991, the project COSMAS (Corpus Search, Management and Analysis System) was launched with the goal of developing an integrated corpus platform and research environment that would enable linguists at the IDS to formulate and refine their queries to the IDS text collections flexibly and independently at their own personal computer. From the beginning, the COSMAS developers subscribed to a set of innovative corpus linguistic methodological principles which are still widely acknowledged in current corpus technology, amongst other things the concept of a unified data model, of multiple concurring linguistic annotations, and most of all the introduction of the concept of a *virtual corpus*. A virtual corpus is a sample of texts drawn from the complete text collection according to a problem-specific view described in terms of external (e.g. bibliographic, specified in terms of metadata) or internal (e.g. the distribution of certain keywords, search patterns or annotations) criteria. Consequently, as of 1992, when the software was first deployed, the COSMAS system offered the tools by means of which users could define their own virtual corpora such that they were representative or balanced w.r.t to their own specific research questions, as well as save or possibly publish them (cf. al Wadi, 1994, p. 132ff).

The successor project COSMAS II was launched in 1995, and from 1996, COSMAS could be used via the internet by linguists all over the world. A part of the project had also been concerned with acquiring more text data, and in 1998, the project *DeReKo I – Deutsches Referenzkorpus* (German reference corpus) started as a cooperation with the universities of Stuttgart and Tübingen. One of its achievements was a mass acquisition of newspaper, fictional, and other text types from publishing houses and individuals, and by the end of the project in 2002, DeReKo contained 1.8 billion tokens. Since then, *Deutsches Referenzkorpus* has been retained as the name of all written corpus holdings at the IDS. By 2004, the IDS corpus extension project had been made a permanent project, and in 2012, DeReKo reached the size of 5 billion word tokens. Since 2008, the IDS has also been a partner in the national and European research infrastructure initiatives TextGrid, D-SPIN, and CLARIN, in which the concept of virtual corpora has been extended and imple-

[†]The authors would like to thank Michael Hanl and and Nils Diewald for their help in preparing the present contribution.

mented to encompass location-independent *virtual collections* (van Uytvanck, 2010).

2. Recent developments

- we organize our acquisition campaigns in waves, addressing 50 to 200 potential license/text donors at a time
- in addition, we approach publishers (in particular the relevant licensing departments) directly at book fairs and sometimes on the phone
- in the negotiations, we seek to acquire licenses as liberal as possible for scientific use, in order of priority: CLARIN-ACA, QAO-NC, QAO-NC-LOC (see Kupietz and Lüngen, 2014)
- chances of convincing rights holders to donate licenses are on average 5%
- the expenses for the acquisition and curation of one word of fictional text are presently around 25,000 times higher than the expenses for one word of newspaper text (see Kupietz, 2014)
- considering only the regularly incoming text data according to existing license agreements, the current growth rate of DeReKo is 1.7 billion words per year

Table 1: Corpus acquisition trivia

As a result of recent campaigns and acquisition deals, DeReKo has grown to over six billion word tokens until 2013, and further grown by a factor of four in the first half of 2014, now containing more than 24 billion word tokens. In the following, we shortly describe the major recent contributions – for more details see Kupietz and Lüngen (2014).

Wikipedia is an example of a web corpus that can be curated under a sufficiently liberal license, and we have made available all German Wikipedia articles and talk pages in DeReKo twice in 2011 (Bubenhofer et al., 2011) and 2013 (Margaretha and Lüngen, in preparation). The 2013 conversion, for example, amounts to more than 1 billion word tokens.
In cooperation with the PolMine project of the University of Duisburg-Essen[1], we have adapted all debate protocols of parliaments in Germany (national and state level) since around the year 2000 (comprising around 360 million word tokens), and we continue to curate the protocols from previous years and other German-speaking parliaments.
We have also curated around 6 million words of fictional text as a result of our 2011 campaign addressing publishers of fiction, while more is in the conversion pipeline.

The bulk of the increase in DeReKo in 2014, however, is formed by a large news database archive for which we obtained DeReKo-specific rights from its commercial provider, containing 102 million documents of press text, specialized journals and specialized e-books. For the latest DeReko release in 2014, we have prepared and included the press data part, containing 98 national and regional newspapers and magazines starting between 2000 and 2003, which amounted to more than 16 billion new word tokens. As a result, the latest DeReKo release contains more than 24 billion word tokens and takes up 550 GB of disk space without annotations (see 2). The new data not only increase the size but also the disperson of genres and topics in DeReKo (see Kupietz and Lüngen, 2014).

- the DeReKo corpus archive (without annotations and version history) uses 550 GB disk space in 1500 files
- for corpus processing, we currently use a machine with 48 cores and 256 GB RAM running CentOS 5.10
- all corpus processing is done in a massively parallel fashion
- for pre-processing of raw data, we use Perl scripts, pdf2html, TagSoup, and tidy
- for the main processing, we use the Saxon Enterprise Edition XSLT-2.0/3.0-Processor
- for coarse quality control, we use Adam Kilgarriff's (2001) measure for corpus similarity
- the POS annotation of the entire DeReKo archive requires between 1 and 2 CPU months for each tool
- the annotation of near-duplicate clusters (mostly carried out within same-source corpora only) (Kupietz, 2005) takes about 7 CPU days
- deriving dependency annotation of the entire DeReKo archive requires between 2 CPU months with Xerox XIP and 13 CPU years with MATE (estimated value, based on a 2.5% DeReKo sample)
- the inter-annotation-tool-agreement on POS tags is typically around 91.5% (see Belica et al., 2011, for details)
- the primary data of DeReKo have been version-controlled and stored in a Subversion repository (currently using 130 GB storage) since 2007
- all DeReKo releases, including primary and annotation data, have been redundantly archived on off-line storage devices since 2009
- long-term preservation and metadata export for OAI-PMH (OAI-PMH, 2008) is currently being migrated to our centre for the long-term preservation of German linguistic research data (Fankhauser et al., 2013)

Table 2: Corpus processing trivia

3. Big Data?

"Big Data" is a broad and fuzzy term, allowing for numerous particular interpretations. Whether it is taken to mean an amorphous mixture or simply an extremely large amount of data of some specific kind, the Deutsches Referenzkorpus DeReKo fulfils both definitions: the latter in a straightforward way, given its current size and growth (see Figure 1), and the former thanks to its status as a primordial sample, from which users can draw virtual corpora (see section 1. and Kupietz et al., 2010).
Figure 2 shows that, measured by the required number of units of the contemporary portable storage medium, the amount of memory needed for the primary text data of DeReKo was actually highest in the beginning in 1969 (factor 400), then decreased to a level of around factor 1 in 1992, where it has remained since then. Only the storage require-

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

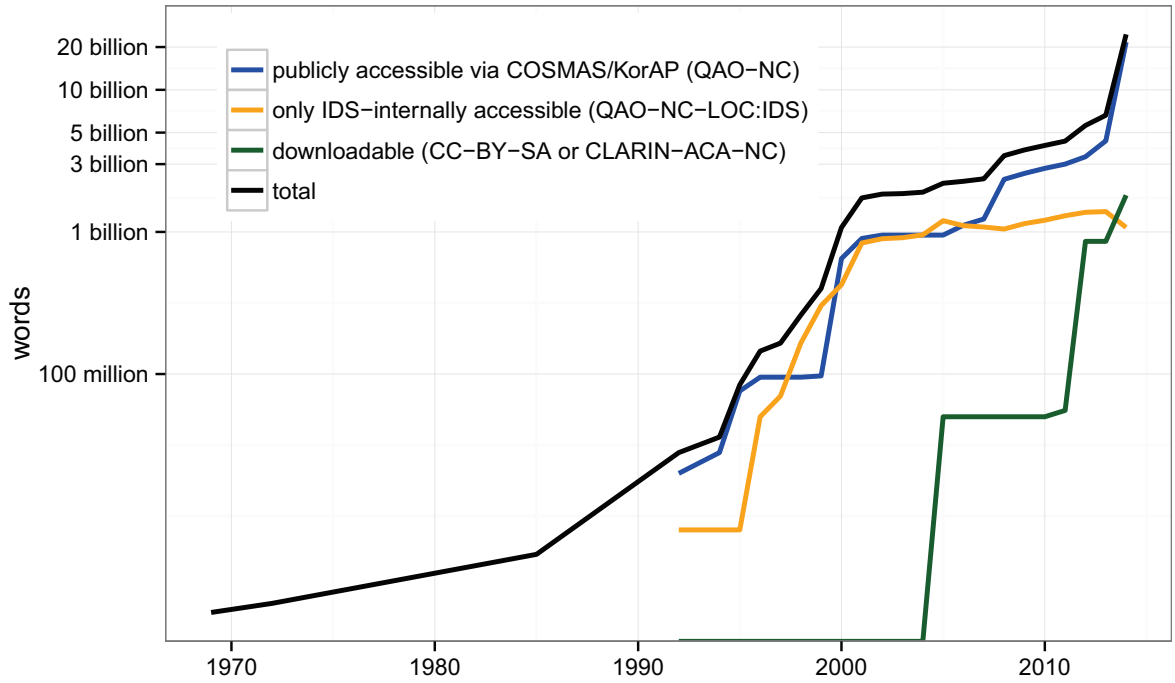

Figure 1: Development of the size of DeReKo in words and its accessibility since 1969 (see Kupietz and Lüngen, 2014, for explanations of the license-abbreviations).

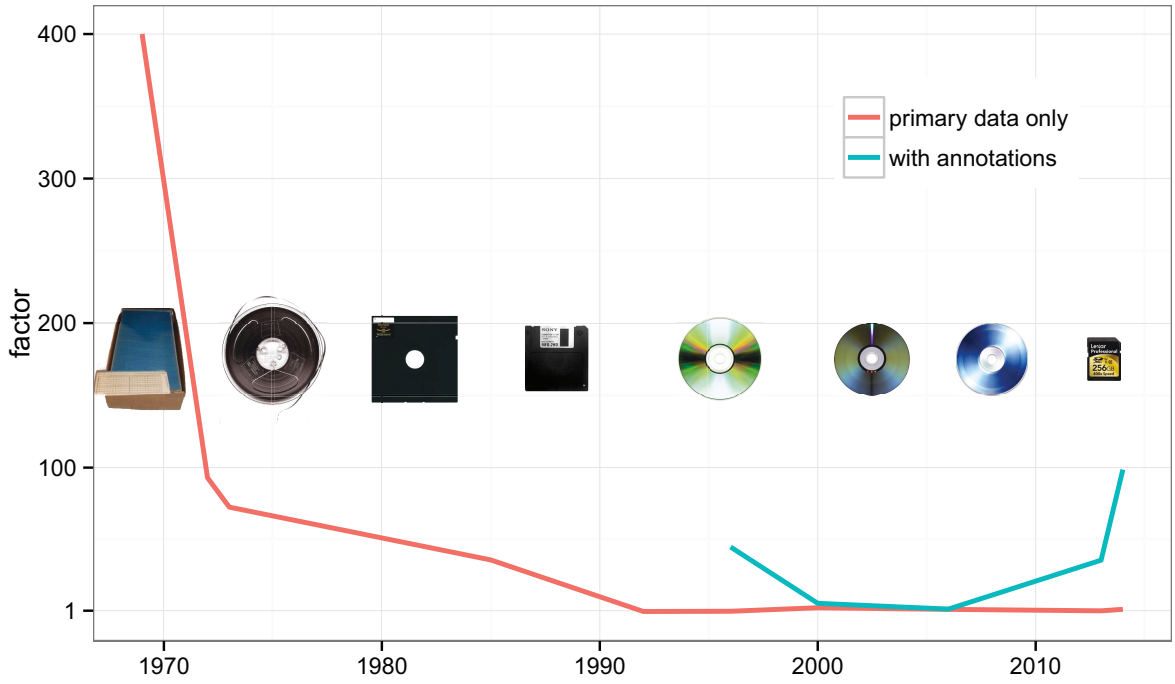

Figure 2: Development of DeReKo's storage requirements since 1969 in relation to contemporary portable storage media. (The figures until 1993 are estimated. For 1969 a box with 2000 punchcards was taken as reference.

ments of the linguistic annotations provided for DeReKo have increased as of 2008, i.e. after the introduction of three new taggers for linguistic annotation (Belica et al., 2011), to a factor of almost 100 of current storage units in 2014 Hence, if we consider the corpus sizes over the years in relation to the available portable storage media, it turns out that the IDS has actually dealt with "big data" ever since the introduction of language corpora.

4. Licensing very large corpora or: "Why don't you just make it open-access?"

Almost as long as its tradition of constructing corpora and making them available to the scientific community is the IDS's tradition of being the notorious bearer of – and sometimes the convenient scapegoat for – the bad news: DeReKo cannot be made available for download. The simple reason

is not that the IDS is a "bad center" that likes sitting on "its data". The reason is of course that the IDS – just like every other provider of comparable corpora – does not have the right to to make the data available for download and that the acquisitions of such a right, e.g. via national licenses, for corpora of the size of DeReKo or a fraction thereof, would hardly be within the limitations of public funding (cf. Kupietz et al., 2010, p. 1849). To our relief, however, this fact is now becoming more and more common ground in the community.[2] This is largely due to the generally increased awareness of intellectual property rights and other personal rights thanks to the Google Books discussion and possibly also thanks to the educational work carried out within CLARIN, often invoking more or less desperate analogies to confront the lack of understanding, such as, e.g., Paweł Kamocki's (2013) analogy with Victor Lustig's selling of the Eiffel Tower in 1925, the image of the car-industry being 'anti-science' due to not providing linguists with free cars, the tightrope that corpus providers walk on (Kupietz, 2009), the cowboy who prefers to shoot first (Ketzan and Kamocki, 2012), or the one with Lady Justice and the balance of interests (next paragraph).

In any case, thanks to this development, approaches aiming at improving the situation for researchers without interfering with the equally legitimate interests of rights holders (on whose donations the researchers vitally depend after all) can nowadays be discussed more openly. As sketched in Figure 3 (Kupietz, 2010), there are more factors involved in such a typical balance of interests and most of them can be gradual. Accordingly, there are many toeholds for such improvements (including the attachment of the scale, in analogy to the legal framework) and to achieve the best results, ideally all of them should be taken into account. One of the very promising approaches is to extend the stack of "technical precautions" in such a way that the possible types of use can be extended to include most types of research that would otherwise only be possible if the corpus were available for download. Part of our current work in this direction is sketched in section 5.2..

Apart from such technical measures along the lines of Jim Gray's (2003) "put the computation near the data" (see also Kupietz et al., 2010; Bański et al., 2012), in our acquisition campaigns, we always try and have always tried to negotiate licenses that are as open as the respective rights holder allows it without the stack of "money" growing too high (see Table 1). Open licenses are great, but what to do if a rights holder does not want to sign them, even if you have been quite inventive to put everything you have on his scale pan?

5. Accessing DeReKo: COSMAS and KorAP

At present, DeReKo is accessible via the Corpus Search, Management and Analysis System COSMAS II (al Wadi, 1994; Bodmer Mory, 2014). It is currently used by more than 32,000 registered users and can handle a DeReKo part of about 7-8 billion words (in one archive) with up to 2 morphosyntactic annotation layers. Due to its having been designed already in the early nineties, its scalability has now reached its limits, because it depends, for example, on holding all indices in RAM and because there are currently no solutions to distribute the system over multiple machines. Because of the above limitations, in 2011 we started the project KorAP (Bański et al., 2012, 2013, 2014), to develop a new corpus analysis platform from scratch, aiming at a scalable and extensible scientific tool, sustainable for a lifecycle of 20 years. Since January 2014, KorAP has been open for IDS-internal alpha testing.

5.1. Scalability

One of the major aims for KorAP has been to achieve horizontal scalability, in order to support a theoretically unlimited number of tokens with a theoretically unlimited number of annotation layers built upon those tokens. This is why KorAP features a multi-component architecture communicating via a simple REST web interface, thus allowing all services to run on different physical machines. The system supports a web UI to allow users to conveniently browse all available data, as well as a policy management component to provide data entry points for restricted resources. These components furthermore include a module to serialize query information (Bański et al., 2014) as well as two search backends that are responsible for processing the query and for retrieving the search results. In order to ensure sustainability, all the components are interchangeable. The backends are based on well-proven Open Source search technologies (Lucene/Solr and Neo4j), which support distributed searches, in this way ensuring horizontal scalability.

5.2. Bringing the computation near the secured data

Another fundamental aim of KorAP was to maximize the research potential on copyright-protected texts. Given the growing amount of textual data and annotations, as well as the growing complexity of the relations among the data components, traditional security measures appear to be inadequate for the task of protecting the integrity of the data while at the same time allowing for fine-grained access control of selected texts and/or selected annotation layers.

KorAP implements a flexible management system to encompass access scenarios that have been identified in the planning phase of the project. That system uses a separate data store to hold security policies for resources that may be subject to restrictions, either imposed by licence agreements concerning the texts and/or the products of annotation tools, or imposed by the users themselves, when they choose to share self-defined virtual collections, potentially containing self-created annotations.

The two backends store DeReKo data in index structures for fast retrieval. Between the web client (the frontend or the API) and KorAP only search requests and results are transmitted. Policy management handles access to different levels of information (raw corpus data, annotation layers, virtual collections, etc.) and authentication based on a ses-

[2] Ironically, however, a slight step backwards was triggered by the open-access movement involving some confusion concerning the actual rights-holders of corpus texts in contrast to the rights-holders of research data in other disciplines, where the rights belong to the researchers themselves or at least to the scientific community.

Figure 3: In the provision and use of corpora, the collision of the basic rights: freedom of science and research, and guarantee of property, in practice, boils down to a balance of interests between researchers and rights holders.

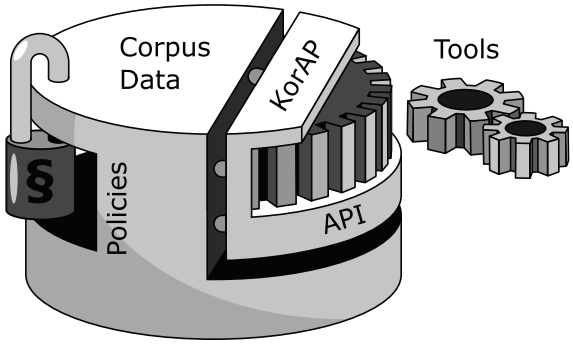

Figure 4: KorAP exposes corpus data for external tools under the control of access policies

sion model (as an intermediate layer between the API entry point and the corpus data).
In addition to standard API access, work is planned on making it possible to create sandboxed environments to support non-remote-API access "near the data", to allow direct access on license protected data and, in general, to provide access to big data without causing unnecessary, if not impossible, network traffic.

6. Conclusions

Written language corpora have been compiled in "big data" dimensions at the IDS since its inception in 1964, and due to the need of curating, storing, annotating, administrating, publishing, and querying them, their continuous expansion has necessarily always been accompanied by pioneering research in the areas of corpus linguistics methodology and corpus technology. The German Reference Corpus DeReKo, as the IDS written corpora collection has been called since the beginning of the millennium, has reached the size of 24 billion word tokens or 30 TB (including linguistic annotations) in 2014, due to the constant influx of text data according to previous license agreements as well as recent acquisitions. Given these dimensions, the scalability capacity of the present corpus management system COSMAS II, which was designed already at the beginning of the 90s, has reached its limits, and a new system called KorAP has been created to ensure continual access to the IDS data.
KorAP aims at horizontal scalability to support an arbitrary number of tokens provided with an arbitrary number of annotation layers. To maximize the research potential on copyright protected texts, it will provide a non-remote API for user-supplied mobile applications.

7. References

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

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)*. European Language Resources Association (ELRA).

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.

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 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.

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.

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.

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

Hellmann, M. W., editor (1984). *Ost-West-Wortschatzvergleiche. Maschinell gestützte Untersuchungen zum Vokabular von Zeitungstexten aus der BRD und der DDR*, volume 48 of *Forschungsberichte des Instituts für deutsche Sprache*. Narr, Tübingen.

Kamocki, P. (2013). Legal Issues: A checklist for data practitioners. Talk given at the EUDAT Workshop on DARUP on 2013-09-23 in Barcelona.

Ketzan, E. and Kamocki, P. (2012). CLARIN-D: Legal and Ethical Issues. Talk given at the Universität des Saarlandes, 2012-03-28.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133. http://www.kilgarriff.co.uk/Publications/2001-K-CompCorpIJCL.pdf.

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. (2009). 45 years of walking the tightrope. Talk given at the D-SPIN/CLARIN-workshop on legal issues 2009-09-21 in Berlin.

Kupietz, M. (2010). Legal and Ethical Issues with respect to LRT and e-infrastructures. Talk given at the D-SPIN Scientific Board Meeting on 2010-12-10 in Berlin.

Kupietz, M. (2014). Der Programmbereich Korpuslinguistik am IDS: Gegenwart und Zunkunft. In Steinle, M. and Berens, F. J., editors, *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, page 298–319. Institut für Deutsche Sprache, Mannheim.

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 (25.5.2010).

Kupietz, M. and Lüngen, H. (2014). Recent Developments in DeReKo. In *Proceedings of LREC 2014*. European Language Resources Association (ELRA).

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

Teubert, W. and Belica, C. (2014). Von der Linguistischen Datenverarbeitung am IDS zur Mannheimer Schule der Korpuslinguistik. In Steinle, M. and Berens, F. J., editors, *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, page 320–328. Institut für Deutsche Sprache, Mannheim.

van Uytvanck, D. (2010). CLARIN Short Guide on Virtual Collections. Technical report, CLARIN. http://www.clarin.eu/files/virtual_collections-CLARIN-ShortGuide.pdf.