The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt

Institute for the German Language (IDS)
R5 6–13, 68161 Mannheim, Germany
{kupietz|belica|keibel|witt}@ids-mannheim.de

Abstract

This paper describes DEREKO (DEUTSCHES REFERENZKORPUS), the *Archive of General Reference Corpora of Contemporary Written German* at the *Institut für Deutsche Sprache* (IDS) in Mannheim, and the rationale behind its development. We discuss its design, its legal background, how to access it, available metadata, linguistic annotation layers, underlying standards, ongoing developments, and aspects of using the archive for empirical linguistic research. The focus of the paper is on the advantages of DEREKO's design as a primordial sample from which virtual corpora can be drawn for the specific purposes of individual studies. Both concepts, *primordial sample* and *virtual corpus* are explained and illustrated in detail. Furthermore, we describe in more detail how DEREKO deals with the fact that all its texts are subject to third parties' intellectual property rights, and how it deals with the issue of replicability, which is particularly challenging given DEREKO's dynamic growth and the possibility to construct from it an open number of virtual corpora.

1. Introduction

Innovative corpus building has a long tradition at the Institute for the German Language (IDS). As early as 1964, Paul Grebe and Ulrich Engel set off the *Mannheimer Korpus 1* project which succeeded in compiling – punchcarding, in fact – a corpus of about 2.2 million running words of written German by 1967. Since that time, a ceaseless stream of electronic text data has been established and fed by a number of subsequent corpus acquisition projects. Today, DEREKO (*Deutsches Referenzkorpus*), the *Archive of General Reference Corpora of Contemporary Written German*, is one of the major resources worldwide for the study of the German language. It currently comprises more than 3.9 billion running words (IDS, 2010) and has a growth rate of approximately 300 million words per year. In compliance with the statutes of the institute that define the *documentation of the German language in its current use* as one of its main objectives, it is a declared IDS policy to provide for the continuous development and long-term sustainability of DEREKO. Some of DEREKO's key features are the following.

- primary purpose: empirical basis for linguistic research
- first created in 1964
- continually expanded
- fictional, scientific and newspaper texts, as well as several other text types
- only complete and unaltered texts (no correction of spelling etc.)
- only licensed material
- not available for download (due to license contracts and intellectual property rights)
- aims to maximize size and stratification
- *primordial sample* design (cf. section 3.)
- allows for the composition of specialized samples
- currently three concurrent annotation layers

2. Access to DEREKO

Not being available for download, the main way to access DEREKO is via COSMAS II, the *Corpus Search, Management and Analysis System* (IDS, 1991-2010). Via this software, DEREKO is currently used by approximately 18,500 registered users. COSMAS II enables its users to compose and analyze so-called *virtual corpora* (see 3.1. and 5.), it provides complex search options (including lemmatization, proximity operators, search across sentence boundaries, logical operators, etc.), it can perform complex (non-contiguous) higher-order collocation analysis (Belica, 1995; Keibel and Belica, 2007), it presents search results in several complementary views, and it provides three alternative interface clients (web client, MS Windows standalone client, script client). A second way to access DEREKO is the Collocation Database CCDB (Belica, 2001-2007; Keibel and Belica, 2007) which serves as the "transparent lab" of the Corpus Linguistics Programme Area at the IDS.

A principal goal is, of course, to make DEREKO available in a way that is suitable for as many applications as possible, while complying with the existing license restrictions (see 2.1.). The options are particularly restricted with respect to applying third-party applications (programs) to DEREKO. One possible approach that we currently entertain in these cases is to invite external researchers to bring or send their programs to us and let them run here – along the lines of: *If the data may not go to the code, let the code go to the data.*

However, as this manual service is rather cumbersome, a more convenient way to bring data and code together is needed: Within the ongoing research infrastructure initiatives CLARIN (Váradi et al., 2008), D-SPIN (Bankhardt, 2009), and TextGrid (Gietz et al., 2006), the IDS currently works on solutions to make DEREKO available in controlled environments through web/grid service application programming interfaces, in a way that respects the legitimate interests of both right holders and linguists.

2.1. Legal background

The reason why DEREKO cannot be made available for download (neither in part nor as a whole) is that the IDS does not own the DEREKO texts. It only has been granted limited rights to use them, and these rights are regulated by more than 150 license contracts with right holders (mostly publishers, in some cases individual authors). The license contracts can be roughly divided into two categories, defined by the eligible user groups: (a) linguists worldwide and (b) IDS employees and guests only.

Restrictions shared by all license contracts specify that (i) only academic use is allowed whereas direct or indirect commercial use is explicitly forbidden; (ii) access is only allowed through specialized software; (iii) only authenticated users may be granted access; (iv) full texts must not be reconstructable from the output of this software; (v) all traffic must be logged; and (vi) abuse must be, as far as possible, prevented by technical precautions. Some of the license restrictions are propagated to the end user by the end user license agreement that every user has to accept before logging on to COSMAS II for the first time.

These rather restrictive license terms are a trade-off between the granted rights of use on the one hand and license costs on the other hand. With less restrictive licenses, a corpus of comparable size would be extremely expensive – if not impossible – to compile.

More generally speaking, as the vast majority of digital research resources in linguistics are subject to third parties' rights, the problem breaks down to a conflict of basic rights, with freedom of science and research on the one hand and the protection of property and general personal rights on the other. As long as the weighting does not shift dramatically in favor of the freedom of science, there will be no general solutions but only compromises, which are more or less specific to individual resource types and research applications. This also means that language resource and service providers are in a delicate middle position between their target communities and the right holders, and that they have to walk the tightrope between the interests of either group. The balance act is particularly risky for both resource providers *and* their communities, because the relationship to right holders is not of a purely legal nature: Both vitally depend on their reputation as trustworthy partners to right holders, whether they are publishing companies or informants.

3. Corpus design

3.1. Ready-to-use vs. primordial samples

Unlike other well-known corpora such as the *British National Corpus* (BNC Consortium, 2007) or the core corpus of the *Digital Dictionary of the 20th Century German Language* (Geyken, 2007), the DEREKO archive itself is not intended to be balanced in any way. The underlying rationale is that the term *balanced* – just as much as the term *representative* – can only be defined with respect to a given statistical population. Therefore, using a pre-existing fixed resource is inefficient as it dictates a specific population to be analyzed. Instead, these issues should, as far as possible, be decided by the individual researcher depending on their general research interests and the specific questions they seek to pursue. For example, it is impossible to state in full generality what specific proportions of text types can be considered balanced or, even more importantly, whether *text type* is a relevant dimension in the first place, or whether it might be more or less relevant than, for instance, the *time* dimension, etc.

It is for these considerations that it was not even attempted to design DEREKO to be balanced, let alone representative. Although the whole archive may be used as a sample itself, the principal purpose of DEREKO is to serve as a versatile *primordial sample* (*"Ur-Stichprobe"*) (cf. Kupietz and Keibel, 2009) from which specialized subsamples, so-called *virtual corpora* (*"virtuelle Korpora"*), can be drawn (cf. section 5.). As a consequence, the further development of DEREKO can focus on the maximization of size and stratification, while the composition of specialized subsamples is left to the usage phase.[1]

Also from a purely practical point of view, the advantages of this approach are numerous. For example, new texts can be continuously included into DEREKO, allowing for research on recent phenomena and changes in language (as in 6.1.). Moreover, virtual corpora can also be maximized for size which is particularly vital when infrequent phenomena are to be investigated. Rare phenomena play a critical role whenever corpus studies go beyond simple frequencies of occurrence: e.g., collocations, grammatical phenomena, frequency analyses across multiple dimensions (e.g., time, topic, and text type). In general, and from an economic point of view, the approach allows for a better exploitation of the available corpus data, as they are reusable for a range of different linguistic research questions that would otherwise require the creation of new corpora from scratch.

To take the economic argument one step further, one may wish to create a virtual corpus from multiple primordial samples which may be distributed across different places. Currently, solutions for such a distributed scenario are being designed and implemented within CLARIN – not only for virtual corpora, but for the more general concept of *virtual collections* (van Uytvanck, 2010) which may contain not only corpus data but any types of language resources.

3.2. Persistency and replicability

Working with a dynamically growing DEREKO archive and especially with a large number of different virtual corpora which are in turn based on numerous different archive states makes it difficult to identify or refer to the specific corpus used in a study. The results of research will therefore not be as easily reproducible as for more static and monolithic corpora. To solve the problem of replicability, all states of the DEREKO archive since the beginning of 2007 are saved, using a standard versioning system. To ad-

[1] In principle, any existing corpus may be used as a primordial sample, but in most cases, it does not serve the actual purpose of a primordial sample very well. Most existing corpora are not both large and stratified enough to allow drawing from them a wide variety of virtual corpora. Only if a corpus is designed as a primordial sample from the outset, its development can focus on maximizing size and stratification.

dress the problem of unique referenceability in particular, a new ISO TC37/SC4 work item proposal for the citation of electronic resources has been submitted (Broeder et al., 2007), which currently has the state of a *draft international standard* (ISO/IEC, 2009). For the case of virtual collections (and thus also for virtual corpora), persistent identifiers will be used that are based on the handle system (Sun et al., 2003).

4. Data and Metadata

4.1. Metadata

One prerequisite for the construction of meaningful samples based on DeReKo is of course its stratification. To this end, knowledge about the basic sampling units (i.e., metadata about single texts or larger units) has to be available. Depending on data source and text type, DeReKo generally provides the following categories of metadata.

Date of publication – for newspaper texts always the day of publication (as provided by the publisher), otherwise sometimes only month or year of publication.

Time period of creation – for newspaper texts always the day of publication (as provided by the publisher); for novels, for example, the period is determined by research or estimation.

Author(s) – available as far as author(s) can be identified. For news agency texts and some newspaper texts unavailable or only available as an abbreviation.

Publisher – always present.

Place of publication – country, city.

Text type – one or multiple items out of a semi-closed set of currently 170 categories (e.g., novel, poem, crime fiction, doctoral dissertation, weather forecast, advertising brochure, horoscope, letter to the editor, guide, etc.)

Topic – the two most likely categories in a two-level taxonomy (top-level: leisure/entertainment, country/society, politics, science, economy/finances, sports, nature/environment, culture, technology/industry, health/nutrition, fiction). Classification is done automatically by means of a naive Bayes classifier (cf. Weiss, 2005).

Near-duplicate properties – any two texts[2] that are covered by more than 70% by common 5-grams are externally (stand-off) linked as *near-duplicates*. The metadata field specifies for each text (i) the number of its near-duplicates, (ii) their maximum, minimum and average similarity value, and (iii) a tentative classification as *copy*, *version*, or *variant* (Kupietz, 2005). The information can, e.g., be used to control biases introduced by reproductions which are in turn due to technical reasons, or to build virtual corpora with a focus on either language production or perception.

Pre-calculated text statistics – number of tokens, words, numbers, sentences, paragraphs, indications of old vs. new orthography, etc.

License conditions – currently one of the following alternatives: copyright-free, GNU Free Documentation License (GFDL), Public, IDS-only, IDS-only-x-parallel-users.

It is worth pointing out – and this is important not only for the construction of virtual corpora –, that the metadata categories differ with respect to their epistemological status. While for example some text statistics are close to plain observations, the topic categorizations are essentially interpretations, and these interpretations do not only depend on the chosen taxonomy, but also on the training data.

4.2. Text model

The DeReKo text model is mainly determined by DeReKo's intended use as a large empirical basis for linguistic research, and additionally restricted by the information that is potentially available or reconstructible at a reasonable cost. This background results in the following key features.

- faithful mapping of content and structure of the source texts
- wide range of supported text types
- hierarchical structure
- annotation of bibliographic, structural and other information necessary or useful for linguistic research and analysis

Currently, the representation format of the text model is an IDS-proprietary, TEI-inspired extension of the Corpus Encoding Standard for XML (XCES) (Ide et al., 2000). As the latest TEI guidelines (P5) (The TEI Consortium, 2007) provide a sufficient degree of adaptability to encode DeReKo without loss of information, a P5-compliant mapping is scheduled for 2010-2011.

4.3. Linguistic Annotation

After DeReKo had already been morphosyntactically annotated twice – with the Logos Tagger in 1995 and with the Gertwol Tagger (Koskenniemi and Haapalainen, 1996) in 1999 –, a new annotation initiative was started in late 2007 with the following fundamental guidelines.

1. Do not rely on judgements of a single tagger. Instead, provide multiple concurring annotations that result from different tools.
2. Use different types of taggers to avoid systematic biases.
3. For each tagger, include as many concurrent interpretations of each linguistic phenomenon as possible.
4. Consider annotating at multiple linguistic levels if appropriate tools are available.
5. Invite an external expert panel to pre-select and recommend tools.
6. After completing the annotation phase, evaluate each annotation layer with respect to fitness for their particular intended use in linguistic research.

As a first result of this endeavor, which is reported in detail in Belica et al. (to appear in 2010), DeReKo-2009-I (IDS,

[2] from the same publisher, or from the same year

2009) was released with three concurrent stand-off annotation layers. These contain part-of-speech and morphological[3] information, provided by the TreeTagger (Schmid, 1994), the Machinese Phrase Tagger (MPT) by Connexor Oy, and the Xerox Finite State Tools (XFST).

For lack of manually annotated test data, we assessed the reliability of the POS tagging results by measuring the agreement between the three tools with respect to single tokens based on a coarse set of nine base tags. The agreement turned out to be the highest for the combination XFST and TreeTagger (95.59%, $\kappa = 0.947$) while TreeTagger and MPT only agreed on 93.47% ($\kappa = 0.921$) of the tokens, XFST and MPT on 93.86% ($\kappa = 0.926$), and all three taggers on 91.57% (Fleiss' $\kappa = 0.931$). Belica et al. (to appear in 2010) provide further details on these evaluations and additionally some thoughts on how to work with automatic annotations, which are by nature not only error-prone but also theory- and implementation-dependent, in the context of empirical linguistic research.

In the meantime, further annotation layers generated by the Xerox Incremental Parser (XIP) suite were added to DeReKo (IDS, 2010). These layers include part-of-speech information, partially disambiguated morphological information, named entities, and syntactic dependencies. Due to license restrictions, however, these XIP layers – as well as the Xerox FST annotation layer – may currently not be made publicly available and have so far only been used for internal testing purposes.

The new annotation layers also lead to some technical challenges, as the sizes of all annotations currently sum up to over 5 TB. Additional annotation layers are planned to be included, among them layers for clause segmentation (e.g., concerning topological fields, Becker and Frank, 2002) and also further layers of part-of-speech annotations with morphological information (e.g., annotations produced by the RFTagger, Schmid and Laws, 2008).

5. Virtual corpora

The general corpus design of a pimordial sample from which specialized virtual corpora are to be drawn is indended to meet the needs of any empirical studies in linguistics, corpus linguistics, and computational linguistics. Potentially, each empirical study may need to compose its own virtual corpus that is specialized to its research questions. Ideally, therefore, this is done by the individual researchers themselves. The most efficient way to define a virtual corpus is in terms of metadata. More precisely, a researcher may ask a corpus query system to randomly draw from a given primordial sample a virtual corpus which displays a set of quantitative and qualitative properties that are pre-specified by the researcher and formulated in terms of metadata (as illustrated in Figure 1). As prerequisites, the following two things need to be available.

- a primordial sample which is sufficiently large and stratified and which provides metadata for the categories that are relevant in the respective study
- an appropriate corpus query system that enables its users (i) to randomly compose virtual corpora by pre-

[3]exception: TreeTagger

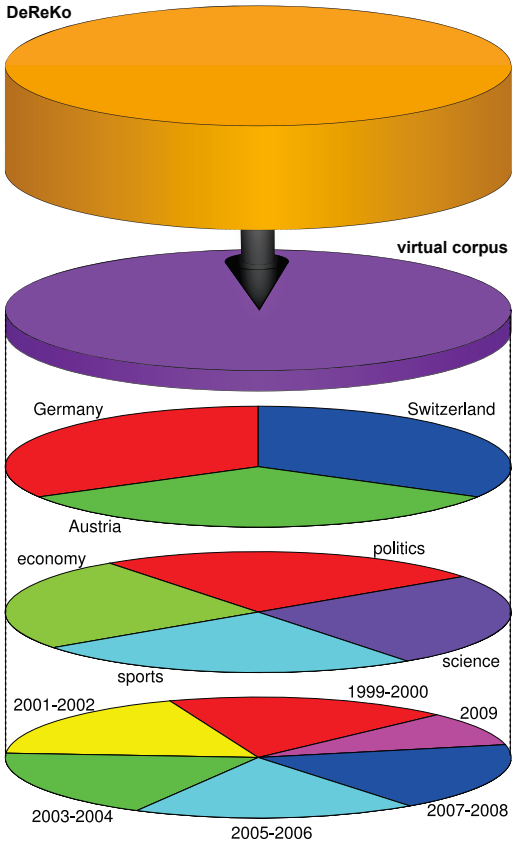

Figure 1: Defining a virtual corpus by specifying its distribution across the metadata dimensions *country of origin* (top), *topic* (center), and *time* (bottom).

specifying their quantitative and qualitative properties in terms of metadata, and (ii) to use these virtual corpora for corpus analyses

To our knowledge, these two prerequisites are presently not fully met for any language. DeReKo itself is fairly close to meeting the first one, but is not yet stratified enough for all conceivable types of research questions. As for the second prerequisite, the associated corpus query system COSMAS II does support the *use* of virtual corpora. However, even though DeReKo is fully annotated with a range of metadata categories (cf. 4.1.), the COSMAS II user interface for *creating* virtual corpora in terms of these metadata are still rather limited. As an interim solution, the DeReKo and COSMAS projects offer to compose virtual corpora by metadata specifications upon request.

6. Application scenarios

As stated above, DeReKo's design as a pimordial sample is indended to meet the needs of any empirical studies in linguistics and related fields. Below, we briefly sketch some application scenarios in which this design displays some of its advantages. In these examples, we assume the two prerequisites listed in section 5. to be fulfilled.

6.1. Investigating language change

Consider a study which seeks to investigate the ways in which a word's collocational behavior changes over time.

One way to approach this would be to compose N different virtual corpora which cover subsequent time periods of equal length and to compare the collocations for the given node word found in each of these corpora (as done by Gondring, 2010). To this end, the researcher may decide to define these virtual corpora by the following qualitative and quantitative properties. First, the individual corpora should have roughly the same size, so the collocations are derived on a comparable statistical basis. Second, each corpus is required to contain only texts produced in the respective time period. Third, in order to ensure that a text's publication date is almost identical to its actual creation date (which is generally not known), the researcher may decide to include only newspaper texts. A second advantage of this decision might be that newspaper texts are available as a virtually continuous stream across time. However, beyond these practical considerations, the researcher should additionally be sure that newspaper texts sufficiently capture the language domain under investigation. Fourth, the researcher may further constrain the target language domain to a certain region (e.g., Eastern Germany) and require to include only texts of newspapers of that region. As a fifth criterion, the researcher may prescribe fixed proportions of text types (e.g., interview, report, news article, comment), and text topics (e.g., politics, economy, sports, science) that all individual corpora are required to display.

6.2. Synchronic language studies

Synchronic studies usually pursue research questions about some contemporary language domain that is defined by a fairly specific point in time (such as *today*). To this end, Belica et al. (2010) propose a general sampling strategy (called the *FReD*[4] *strategy*) for building a corresponding synchronic corpus. It does this by specifying the relative number of texts to be included in the corpus from each past time slice. While the general FReD strategy has to be fine-tuned to the specific language domain, it usually ascribes more recent time slices a greater sampling weight (as in Figure 2).

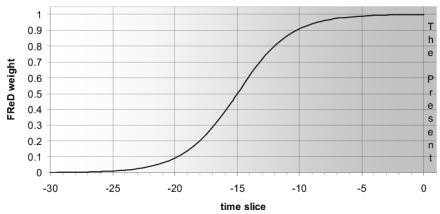

Figure 2: A specific function of FReD sampling weights.

With the primordial sample design of DEREKO synchronic virtual corpora may be defined in a straightforward way. After the researcher has fine-tuned the general FReD strategy to the specific language domain they wish to investigate, they may ask the corpus query system to randomly draw from DEREKO a virtual corpus whose distribution across time matches these FReD sampling weights. Of course, the researcher will usually want to combine this sampling strategy for time with sampling strategies for other dimensions that appear relevant for the given language domain (cf. above).

[4] *Frequency Relevance Decay*

6.3. Other scenarios

There is a large variety of other research scenarios which benefit from DEREKO's primordial sample design. For instance, a study may be interested in contrasting language use between regional varieties of written German (e.g., Austria vs. Germany vs. Switzerland). Suitable virtual corpora could be defined by requiring them to contain only data from the respective region and to be comparable in all other dimensions (e.g., identical sizes, roughly the same distribution of text types, etc.).

As another example, a researcher may be only interested in a specific phenomenon (and not so much in a specific domain) and decide to build a virtual corpus containing all – or a random sample of all – instances of this phenomenon to be found in DEREKO (or in some pre-existing virtual corpus). Note that such a virtual corpus cannot be defined in terms of text-external metadata (like those listed in 4.1.) but rather in terms of text-internal properties which do not describe the text as a whole but only parts of it. This essentially corresponds to an ordinary corpus query of the form "find/extract all texts containing *X*" where the phenomenon *X* may be specified in terms of surface forms and annotations.

7. Discussion

This paper described DEREKO from various angles: corpus design, underlying standards, access, legal aspects as well as available metadata and annotation layers. Being the only major corpus archive designed as a primordial sample, special attention was given to how DEREKO may best serve the needs of actual corpus studies. As the investigated language domains and research questions differ between studies, practically every individual study is advised to consider creating a virtual corpus from DEREKO that is specialized to its specific research interests.

The full potential of this proposal is reached when both the primordial sample and the associated corpus query system fulfill certain prerequisites (cf. section 5.). For the case of DEREKO and COSMAS II, these prerequisites are not yet fully met, and it is a long-term IDS goal to close the gap.

In particular, one current focus of the DEREKO project is to increase the degree of stratification by acquiring textual data for registers and text types that are currently not well represented in the archive (e.g., fiction and academic texts). As for the variety of German newspapers and magazines, the project appears to approach a ceiling: After having contacted most larger ones (both in- and outside Germany), we by now have acquired texts from most of them that are in principle both willing and able to support DEREKO.

Other directions for future research include the following. First, the technical aspects to realize the notion of *virtual collections* (cf. 3.1.) and the issue of persistent identification and re-usability (cf. 3.2.) will be developed at the level of the CLARIN initiative.

Second, we plan to make available various pre-defined virtual corpora with frequently requested properties.

Third, as DEREKO itself may not be offered for download (cf. 2.1.), we currently look into the feasibility of upgrading the license agreements in such a way that they allow us (i) to make parts of DEREKO available within e-infrastructures like CLARIN, and (ii) to publish a scrambled version of parts of DEREKO. Such a scrambled corpus is constructed by randomly shuffling complete sentences – thus, it still allows linguistic analyses at the sentence level while preventing that any sequences greater than sentences can be reconstructed.

8. References

Christina Bankhardt. 2009. D-Spin – Eine Infrastruktur für deutsche Sprachressourcen. *Sprachreport*, 1/2009:30–31.

Markus Becker and Anette Frank. 2002. A stochastic topological parser of German. In *Proceedings of COLING 2002. Taipei, Taiwan, Province of China*, pages 71–77.

Cyril Belica, Holger Keibel, Marc Kupietz, Rainer Perkuhn, and Marie Vachková. 2010. Putting corpora into perspective: Rethinking synchronicity in corpus linguistics. In *Proceedings of the 5th Corpus Linguistics Conference (CL 2009)*, Liverpool. University of Liverpool.

Cyril Belica, Marc Kupietz, Harald Lüngen, and Andreas Witt. to appear in 2010. The morphosyntactic annotation of DEREKO: Interpretation, opportunities and pitfalls. In M. Konopka, J. Kubczak, C. Mair, F. Šticha, and U. Wassner, editors, *Selected contributions from the conference Grammar and Corpora 2009*, Tübingen. Gunter Narr Verlag.

Cyril Belica. 1995. Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache.

Cyril Belica. 2001-2007. Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Mannheim: Institut für Deutsche Sprache. `http://corpora.ids-mannheim.de/ccdb/`.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. `http://www.natcorp.ox.ac.uk/`.

Daan Broeder, Thierry Declerck, Marc Kemps-Snijders, Holger Keibel, Marc Kupietz, Lothar Lemnitzer, Andreas Witt, and Peter Wittenburg. 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`.

Alexander Geyken. 2007. The DWDS corpus: A reference corpus for the German language of the twentieth century. In Christiane Fellbaum, editor, *Idioms and collocations: Corpus-based linguistic and lexicographic studies*. Continuum, London.

Peter Gietz, Andreas Aschenbrenner, Stefan Büdenbender, Fotis Jannidis, Marc Wilhelm Küster, Christoph Ludwig, Wolfgang Pempe, Thorsten Vitt, Werner Wegstein, and Andrea Zielinski. 2006. Textgrid and ehumanities. In *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing E-SCIENCE ´06*, Amsterdam. IEEE Computer Society 2006.

Oliver Gondring. 2010. Diachroner Wandel von Kollokationsprofilen: Zur Emergenz sprachlicher Regelhaftigkeit am Beispiel von Neologismen. Gastvortrag am 14.01.2010, Gesprächsrunde, Institut für Deutsche Sprache, Mannheim.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 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. 1991-2010. COSMAS I/II Corpus Search, Management and Analysis System. Mannheim: Institut für Deutsche Sprache. `http://www.ids-mannheim.de/cosmas2/`.

IDS. 2009. Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2009-I (Release vom 28.02.2009). Mannheim: Institut für Deutsche Sprache. `http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html`.

IDS. 2010. Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010-I (Release vom 02.03.2010). Mannheim: Institut für Deutsche Sprache. `http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html`.

ISO/IEC. 2009. ISO/DIS 24619: Language resource management – persistent identification and access in language technology applications. Technical report, International Organization for Standardization, Geneva, Switzerland, 4. September.

Holger Keibel and Cyril Belica. 2007. CCDB: A corpus-linguistic research and development workbench. In *Proceedings of the 4th Corpus Linguistics Conference (CL 2007)*, Birmingham. University of Birmingham. `http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf`.

Kimmo Koskenniemi and Mariikka Haapalainen. 1996. GERTWOL – Lingsoft Oy. In Roland Hausser, editor, *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*, chapter 11, pages 121–140. Niemeyer, Tübingen.

Marc Kupietz and Holger Keibel. 2009. The Mannheim German Reference Corpus (DEREKO) as a basis for empirical linguistic research. In Makoto Minegishi and Yuji Kawaguchi, editors, *Working Papers in Corpus-based Linguistics and Language Education*, volume 3. Tokyo University of Foreign Studies (TUFS), Tokyo.

Marc Kupietz. 2005. Near-duplicate detection in the IDS corpora of written German. Technical Report kt-2006-01, Institut für Deutsche Sprache, Mannheim. `ftp://ftp.ids-mannheim.de/kt/`

ids-kt-2006-01.pdf.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester.

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

S. Sun, L. Lannom, and B. Boesch. 2003. *Handle System Overview*. Number 3650 in Request for Comments. IETF, November. http://www.ietf.org/rfc/rfc3650.txt.

The TEI Consortium, editor. 2007. *Guidelines for Electronic Text Encoding and Interchange (TEI P5)*. The TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

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

T. Váradi, P. Wittenburg, S. Krauwer, M. Wynne, and K. Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. In *Proceedings of the 6th International Language Resources and Evaluation Conference (LREC'08)*, Paris. European Language Resources Association (ELRA).

Christian Weiss. 2005. Die thematische Erschließung von Sprachkorpora. Technical report, Institut für Deutsche Sprache, Mannheim. http://www.ids-mannheim.de/kl/projekte/methoden/te/te.pdf.