『コーパスに基づく言語学教育研究報告』No.3（2009）

The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research

Marc Kupietz and Holger Keibel
Institute for the German Language (IDS), Mannheim

Abstract

This paper describes DEREKO, the archive of general reference corpora of contemporary written German at the IDS Mannheim, and the rationale behind its development. We discuss its corpus design, its current composition, issues about its further development, available metadata and aspects of using the archive for empirical linguistic research.

1. Introduction

Corpus building has a long tradition at the Institute for the German Language (IDS). As early as 1964, Paul Grebe und Ulrich Engel set off the *Mannheimer Korpus 1* project which succeeded in compiling (punchcarding, actually) 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 at the IDS, is one of the major resources worldwide for the study of the German language. It currently comprises 3.4 billion words and has a growth rate of approximately 300 million words per year (see Figure 1). In compliance with the statutes of the institute as a public-law foundation that define the '*documentation of the German language in its current use'* as one of its main goals, it is a declared IDS policy to provide for a long term sustainability of DEREKO. At present, the standing project KAB *(Ausbau und Pflege der Korpora geschriebener Gegenwartssprache)* is responsible for further DEREKO development and maintenance.

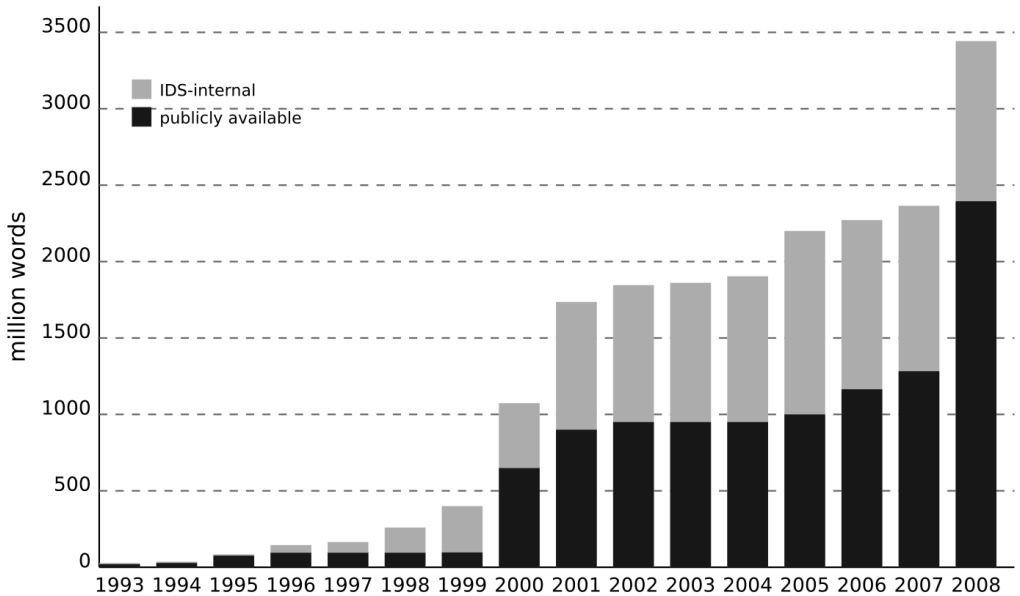

Figure 1: Development of the size of the DEREKO archive since 1993

Some of the key features of DEREKO are the following:

- largest linguistically motivated collection of German texts
- developed since 1964
- established in 1964
- continually expanded
- contains texts starting from 1956
- contains fictional, scientific and newspaper texts, as well as several other types of text
- contains only complete texts
- contains only unaltered texts (no correction of spelling etc.)
- contains only licenced texts

The main purpose of DEREKO is to serve as an *empirical basis for the scientific study of contemporary written German*. It is in general most useful whenever the focus of interest is on the language itself rather than on the information conveyed by it. DEREKO is not intended for the use in research areas like computational linguistics, information retrieval or language technology in general. Apart from the IDS' focus on its target community (viz. German linguistics) this is also because of legal (namely IPR) restrictions that do not permit the institute to offer full texts for download.

2. Using DeReKo

Not being available for download, the main way to access DEREKO is via *COSMAS II,* the *Corpus Search, Management and Analysis System*. The COSMAS II software is developed in a separate project, situated in the computing centre of the IDS. Via this software, DEREKO is currently used by approximately 15,000 registered users. COSMAS II allows for the composition of so-called virtual corpora, it provides complex search options (including, e.g., lemmatization, proximity operators, search across sentence boundaries, logical operators), can perform complex (non-contiguous) higher order collocation analysis, features various views for search results and different interface clients (Web, MS Windows software-client, script-client). A second way to access DEREKO is the Collocation Database CCDB which serves as the "transparent lab" of the Corpus Linguistics Programme Area (cf. Keibel/Belica 2007).

In the course of language resource infrastructure initiatives like CLARIN[1] and other *eHumanities* projects like for instance TextGrid[2], DEREKO will also be made available through web service application programming interfaces.

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 is not the owner of the DEREKO texts. Rather, it only has been granted limited rights to use them. These rights are regulated by over 150 licence contracts with copyright holders (mostly publishers, also in some cases individual authors). The licence contracts can be roughly divided into two categories in their eligible users groups: linguists in general and staff members of the IDS only. Restrictions shared by all licence 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) that 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 licence restrictions are propagated to the end user by the *end user licence agreement* that every user has to sign before using COSMAS II the first time. To retain its long-time outstanding reputation as a trustworthy partner of text donors in Germany and abroad, the IDS maintains the highest standards of ethical and meticulous legal conduct pertaining to the DeReKo use and encourages and expects its employees to report any suspected violations of the *end user licence agreement*.

[1] *Common Language Resources and Technology Infrastructure. Network:* http://www.clarin.eu/
[2] *TextGrid - a modular platform for collaborative textual editing*. http://www.textgrid.de/

3. Corpus design

3.1 Ready-to-use vs. primordial samples

Unlike other well-known corpora, like, e.g. the *British National Corpus (BNC)* or the core corpus of the *Digital Dictionary of the 20th Century German Language*[3], the DEREKO archive itself does not intend to be *balanced* in any way. The rationale behind this is that the term *balanced* – just as much as the term *representative* – can only be defined with respect to some given statistical population. The resource itself should not dictate a specific population, nor should it define which properties of the population are of particular relevance. Instead, these issues should, as far as possible, be decided by the individual researcher depending on their general research interests and the specific question they seek to answer. For example, it is impossible to state in full generality what specific proportion of text types can be considered balanced or, even more importantly, that *text type* might not be a relevant dimension in the first place, or that it might be less relevant than, for instance, time, 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 superordinate sample, or *primordial sample (Ur-Stichprobe)*, from which specialized subsamples, so-called *virtual corpora (virtuelle Korpora)* can be drawn. This means that the design and further expansion of DEREKO can focus on the maximization of size and stratification, while the composition of specialized samples is left to the usage phase.

3.2 Building samples in COSMAS II

In COSMAS II, DEREKO-based samples can currently only be constructed manually, not on a text basis, but on the basis of logical internal units such as a set of books or an entire month of editions from the same newspaper. To this end, improvements are under way and already prototypically implemented. The goal of these developments is to allow users to build virtual corpora by any of the following criteria or procedures, respectively.

- by an explicit text reference within the source archive
- by specifying the desired frequency distribution of any metadata within the target corpus
- by "upgrading" any hit set obtained by some corpus query to a new personal corpus

3.3 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 leads to the problem that the respective bases of empirical linguistic research are difficult to identify or to refer

[3] *Das digitale Wörterbuch der Deutschen Sprache des 20. Jh. (DWDS)*. http://www.dwds.de/

to such that the results of research will not be as easily reproducible as in the times of static monolithic corpora like the BNC. To solve the problem of replicability, all states of the DEREKO archive are saved since the beginning of 2007, using a common versioning system. To address 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, ISO24619).

4. Data and Metadata

4.1 Metadata

One prerequisite for the construction of meaningful samples based on DEREKO is of course the presence of explicit knowledge (or knowledge that can be made explicit) about the basic units of such a construction. In other words, *metadata* about single texts or larger units have to be available. What kinds of metadata are currently available in DEREKO depends on the data source and the text type. In general the following categories are available:

- date of publication
- time period of creation
- name of author
- name of publisher
- medium of publication (newspaper, book, news agency, …)
- place of publication
- text type (partially)
- topic category
- information about (near-)duplicates
- size of near-duplicate class
- number of words/sentences/paragraphs
- indications of old orthography vs. new orthography
- licence condition (available inside/outside the IDS)

These metadata – or rather categories of metadata – differ qualitatively and can be classified along several dimensions, like for example *type* (descriptive, structural, administrative, legal), *source* (publisher, probabilistic computer programs, human experts), etc. Along these dimensions, some *meta-metadata* would be desirable, especially concerning the formal confidence with respect to the metadata. Currently, such meta-metadata are only available in the form of documentation.

4.2 Text model

The DEREKO text model is mainly determined by its intended use, namely as a *non-downloadable empirical basis for linguistic research*, and additionally restricted by the information that is potentially available. This background results in the following key features:

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

The pivot representation of the text model currently is an IDS-proprietary format termed BOT. This format can, without loss, be mapped to proprietary, TEI-inspired, extensions of CES and XCES (IDS-XCES) and also to standard XCES. A mapping compliant to the TEI-P5-guidelines (TEI Consortium, 2008) and also a corresponding migration of the software infrastructure are planned.

4.3 Conversion of raw texts

The raw texts come in all possible formats and unsystematic variations. For this reason it would be very inefficient to develop a complete specialized converter for every format variant. Instead a funnel-like dataflow architecture is applied, so that small bits of software (developed in Perl and XSLT 2.0), increasingly general from top to bottom, can push the dataflow into very few intermediate formats and finally into an internal representation conforming the DEREKO text model from where different outlets lead to the different concrete output representations. The quality of the conversions is controlled manually (on the basis of random samples) and automatically (by comparisons with reference samples). These comparisons are based on distributions of characters, words, and 5-grams (see Kupietz 2005; Belica et al. 2007).

4.4 Linguistic Annotation

Currently, apart from sentence boundaries and functional lemmatization, only some small parts of DEREKO are linguistically annotated. For 2009, however, a full multiple and concurrent annotation at several linguistic levels (at least part of speech and syntactic relations) is planned for the entire DEREKO archive. The prime argument for offering multiple and concurring annotations is that the annotation has not the status of *observed data*, instead it constitutes a theory- and implementation-dependent interpretation. In addition, because of the size of DEREKO, neither manual annotation nor manual control of the automatic annotation would be feasible. In consequence, the expected degree of inaccuracy is very high, particularly wherever, for example, lexical or grammatical variation, i.e., linguistically challenging phenomena are concerned. Given

these facts, a researcher exploiting only a single type of annotation will run the risk of not actually studying the language, but rather the annotation tool or the theory behind it. Uncontrollable and possibly systematic type II errors (false negatives) that can hardly be avoided when using annotation based searches further add to this risk. The only viable approach to alleviate – albeit not to solve – this problem is to use a range of tools that are based on different theories. However, linguistic analyses can of course also be conducted without using linguistic annotations at all.

References

Belica, Cyril / Keibel, Holger / Kupietz, Marc / Perkuhn, Rainer. (2007) "Web as Corpus: Kooperation mit der Universität Bologna", *Sprachreport Sonderheft/März 2007. Auslandskooperationen des Instituts für Deutsche Sprache*, 21-25.

Broeder, Dan / Declerck, Thierry / Kemps-Snijders, Marc / Keibel, Holger / Kupietz, Marc / Lemnitzer, Lothar / Witt, Andreas / Wittenburg, Peter. (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

Institut für Deutsche Sprache (1991-2008) *COSMAS I/II (Corpus Search, Management and Analysis System)*. http://www.ids-mannheim.de/cosmas2/

ISO CD 24619 (2008-02-08) *Citation of Electronic Resources*, International Organization for Standardization, Geneva, Switzerland.

Keibel, Holger / Belica, Cyril. (2007) "CCDB: A Corpus-Linguistic Research and Development Workbench", *Proceedings of Corpus Linguistics 2007*, Birmingham. http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf

Kupietz, Marc. (2005) *Near-Duplicate Detection in the IDS Corpora of Written German* (Tech. Rep. KT-2006-01), Institut für Deutsche Sprache.

Kupietz, Marc. (2008): "DeReKo durchbricht Drei-Milliarden-Grenze", *Sprachreport* 2/2008: 28-Mannheim: Institut für Deutsche Sprache.

TEI Consortium. (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/