Publsihed in: Arranz, Victoria/Eerten, Laura van (eds.): Language Resources: From Storyboard to Sustainability and LR Lifecycle Management, Workshop held at the seventh conference on International Language Resources and Evaluation (LREC 2010). Malta. - European Language Resources Association (ELRA), 2010. pp. 39-43.

# The German Reference Corpus: New developments building on almost 50 years of experience

**Marc Kupietz, Oliver Schonefeld, Andreas Witt**

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

**Abstract**

This paper describes the efforts in the field of sustainability of the *Institut für Deutsche Sprache* (IDS) in Mannheim with respect to DEREKO (Deutsches Referenzkorpus) the *Archive of General Reference Corpora of Contemporary Written German*. With focus on re-usability and sustainability, we discuss its history and our future plans. We describe legal challenges related to the creation of a large and sustainable resource; sketch out the pipeline used to convert raw texts to the final corpus format and outline migration plans to TEI P5. Due to the fact, that the current version of the corpus management and query system is pushed towards its limits, we discuss the requirements for a new version which will be able to handle current and future DEREKO releases. Furthermore, we outline the institute's plans in the field of digital preservation.

## 1. Introduction

The Institute for the German Language (IDS) has a long tradition in building corpora. DEREKO (*Deutsches Referenzkorpus*), the *Archive of General Reference Corpora of Contemporary Written German*, has been set off as the Mannheimer Korpus 1 project in 1964. Paul Grebe and Ulrich Engel succeeded in compiling a corpus of about 2.2 million running words of written German by 1967. Since then, further corpus acquisition projects established a ceaseless stream of electronic text documents and let the corpus to grow steadily (Kupietz & Keibel, 2009).

As of 2010 the corpus, which is intended to serve as an empirical basis for Germanic linguistic research, compromises more than 3.9 billion words (IDS, 2010) and has a growth rate of approximately 300 million words per year. 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 declared IDS policy to provide for a long term sustainability of DEREKO. In 2004 a permanent project responsible for its maintenance and further development has been established.

## 2. Current state

As stated in Kupietz et al. (2010), the key features of DEREKO are the following:

- established and developed in 1964
- contains texts from 1956 onwards
- continually expanded
- contains fictional, scientific and newspaper texts as well as several other types of text
- 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)
- maximum size, primordial sample design
- allows the composition of specialized samples
- endowed with currently three concurrent annotation layers

Unlike other well-known corpora like, e.g. the BNC (BNC, 2007), DEREKO 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 specific research question and some statistical population. Thus the composition of a sample should be part of the usage phase and not part of the design phase of a corpus that shall be used as a general basis for empirical linguistic research. As a consequence of this so called *primordial sample* approach, the text acquisition can concentrate on the maximization of size and stratification and as any DEREKO-based samples can be defined an overall boost of versatility and re-usability is achieved. A more detailed view of DEREKO's primordial sample approach and its application scenarios is given in Kupietz et al. (2010).

### 2.1. Legal aspects of re-usability

To allow for a broad sampling of language data, the IDS has negotiated license contracts with various copyright owners, such as authors, publishing houses and newspapers. The contracts grant non-commercial academic use of the data exclusively and allow access only via software that among other things must prevent the reconstructability of whole texts. Licenses are open-ended, but can be cancelled by the licensor at any time. As a consequence with respect to sustainability, the IDS cannot guarantee the persistency of texts contained in DEREKO as the right holders can in principle withdraw the right of use any of their texts at any time. In the last years, however, this happened only to single newspaper texts. The most frequent reason was that a publisher had undertaken to refrain from the further distribution of an article. As the average frequency of such deletions was less than 50 per year, until now, the replicability of DEREKO-based findings should not have been significantly affected.

At large, the situation concerning usage rights and their sustainability ist not ideal, but like all large-scale corpus

projects, DeReKo, more specifically the IDS as the language resource and service provider, has to walk a tightrope between the interests of its target community and those of the IPR holders. More generally speaking, as the vast majority of digital research resources in linguistics are subject to third parties' rights, the problem boils 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.

The IDS is involved in campaigns for a more research friendly copyright-law, e.g. via the Leibniz Association and in CLARIN. In the context of CLARIN and the German counter-part D-SPIN, the IDS also works on improved licensing models. One approach we follow for example in the context of CLARIN and D-SPIN is to develop upgrade agreement models with a graded transferability of usage rights and to test them with selected licensors of DeReKo-texts in order to improve their re-usability within secure distributed research infrastructures.

## 3. Annotations

In 1993, the IDS started COSMAS II (IDS, 1991–2009), the Corpus Search, Management and Analysis System, as a first step towards providing an access to linguistic annotations. It was planned in order to specifically be capable of handling multi-layer annotations. In 1995 DeReKo was enriched with annotations from the Logos Tagger and in 1999 the analysis from Gertwol Tagger were added.

The IDS recently has started an extensive corpus annotation venture to provide even more annotations. As described in Belica et al. (to appear in 2010), Machinese Phrase Tagger from Connexor Oy, the TreeTagger from Stuttgart University (Schmid, 1994) and the Xerox FST Linguistic Suite and various custom filters have been applied on DeReKo to produce concurrent stand-off annotations. In a first step only the morphological and the part-of-speech analysis components were considered. This annotation process took about 6 CPU-years and resulted in about 3.5 TB of data. In the meantime, DeReKo was also annotated on the syntactic level with the Xerox Incremental Parser XIP. Currently, however, the IDS has only acquired sufficient licenses to make TreeTagger and Connexor annotations available to the outside world via COSMAS II. Presumably because of the danger of reverse engineering, that would arise when a large annotated corpus was made publicly accessible without restrictions, the problems of for acquiring sufficient licenses for commercial taggers and parsers are comparable to those for copyrighted text.

DeReKo-2009-I (IDS, 2009) was the first release with annotations. These contain part-of-speech and morphological (except TreeTagger) information, provided by the above mentioned tools. A detailed report on the annotation process, an assessment of their reliability, and some thoughts on how to use them methodologically sound in linguistic research can be found in Belica et al. (to appear in 2010).

## 4. Re-usability and sustainability

### 4.1. From raw data to corpus representation formalisms

The stream of raw data that constantly feeds DeReKo with currently about one million words per day is supplied by the text donors in many different formats. Mostly, these formats are tailored towards the requirements of the publishing industry. However, for the purpose of analysing the data, it has to be converted to a common format. The IDS has developed a format based on XCES (Ide at al., 2000). The input data is converted through a pipeline of various transformation steps. While due to its funnel-like architecture with many small specialized filters only at the beginning of the processing pipeline, a large part of this transformation system is re-usable also for new data sources, the process is still quite an expensive task because often manual intervention is needed due to the broad variances in the input, even for data coming from a single source. Figure 1 gives an overview of the whole processing pipeline.

Recently, the IDS has started to investigate a migration of DeReKo from the custom XCES variant to TEI. As the TEI P5 guidelines (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. Besides the obvious advantages of a most recent version of the standard such a conversion does also have drawbacks: Parts of the processing pipeline as well as a large portion of the quality assurance battery are tailored to the old format and migrating to TEI P5 would not gain an immediate advantage. Since DeReKo is not available for direct download, no one outside the IDS will directly benefit from this conversion. In addition there are currently no tools for processing TEI-P5-compliant data that we know of which could be applied on DeReKo. The vast amount data is also beyond the editing and validation capabilities of any current XML editor. For now, the main immediate advantages concerning interoperability, though also only IDS-internally, will arise from the migration from DTD based to schema based validation, which allows for a finer grained control of data types and better maintainability. In the long run, however, we hope that with a migration to TEI we will contribute to a harmonization and standardization process which after all will also lead to tools that are able to deal with large scale TEI data.

Furthermore, migrating to TEI will save us the reinvention of the wheel for areas that are not yet fully covered by the IDS-XCES formalism. For example, TEI offers the opportunity to exploit the standardized feature structures to describe different annotation layers in a unified representation. Witt et al. (2009) gives a detailed view on how to adopt feature structures to archive this goal and discusses advantages and disadvantages of this approach.

### 4.2. Persistence and preservation

Unlike more static or monolithic corpora, DeReKo being constantly improved and expanded, also has to deal with challenges in the context of replicability of DeReKo-based research, data persistence, and persistent reference. To ensure that all data states are, in principle, reproducible DeReKo is maintained in a subversion repository since

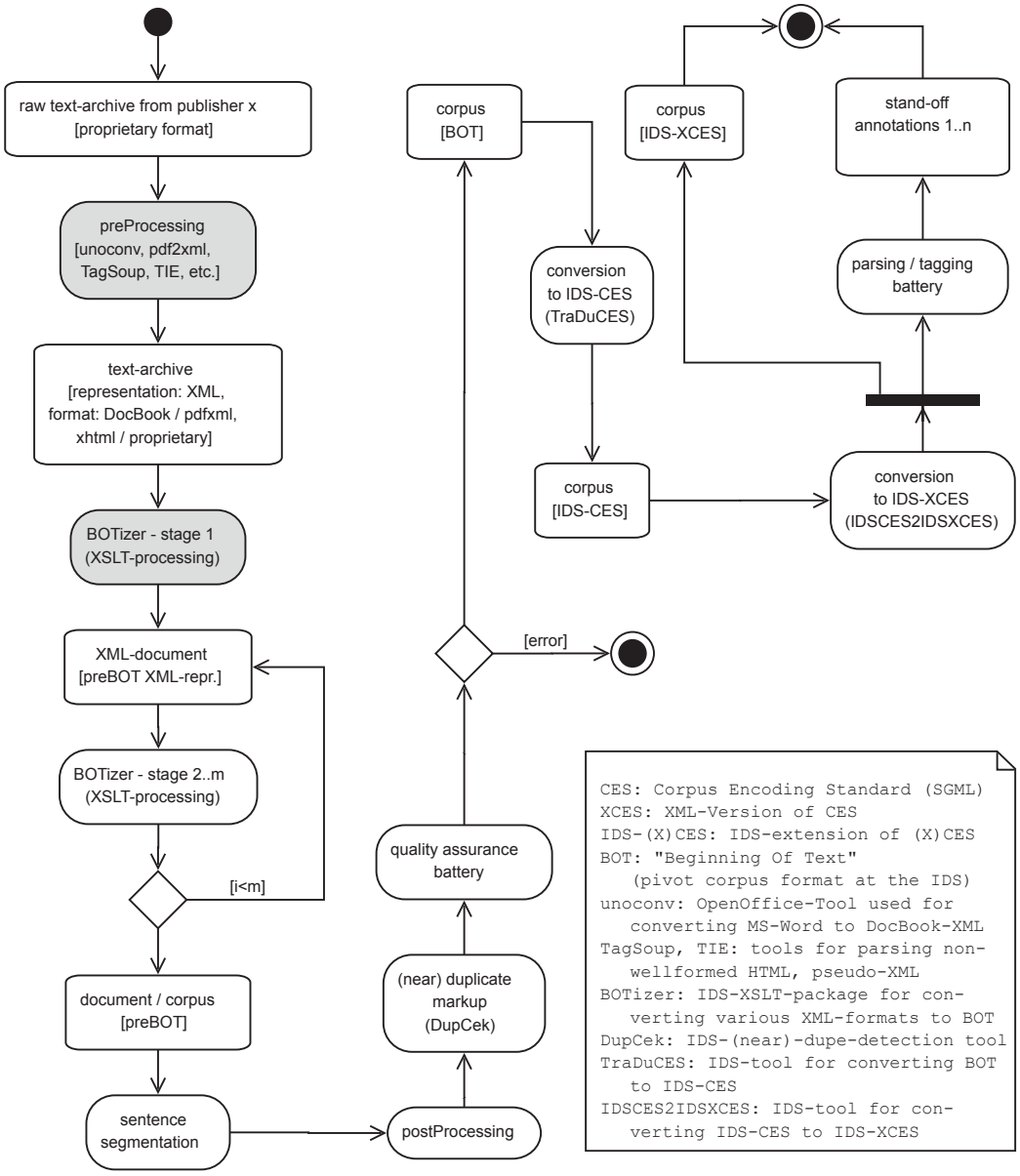


Figure 1: Architecture for processing raw texts. The filter steps highlighted in gray are decreasingly dependent on the input format. Most of the architecture can be reused for new formats. For the migration to TEI P5, first a converter from IDS-XCES will be implemented for testing purposes and evaluation. For a complete migration the following steps will be necessary: (i) insertion of a new conversion routine from preBOT to TEI before the sentence segmentation, (ii) adaption of subsequent steps (quality assurance battery, etc.), (iii) removal of IDS-CES- and IDS-XCES-conversion.

the beginning of 2007. However, with this approach taken alone, the reproduction of old states so that they are actually usable is expensive because a complete version of DEREKO has currently a size of about 5 TB and to make it usable via COSMAS requires at least partial re-indexing. A possible solution to this problem could be to integrate versioning into the core database system. We will consider this in the development of a new corpus search and analysis system (see following section).

To be able to persistently refer to corpora, documents, and texts contained in different DEREKO archive states, internally unique persistent identifiers are used. In the context of the CLARIN initiative, we are currently planning to combine these with globally unique identifiers based on the handle system (Sun et al. 2007), for example to allow for the construction of distributed virtual corpora or resource collections (cf. Kupietz et al. 2010). Together with the ISO standard for the persistent identification of electronic language resources (cf. Broeder et al., 2007; ISO/DIS 24619: ISO/IEC, 2009) this will allow for accurate reference to and citation of DEREKO or parts of it and ensures the traceability of DEREKO-based research.

To further secure the sustainability of DEREKO, the IDS is currently working on a digital preservation strategy. Especially the current the legal arrangements pose a problem for an off-site archiving of the resources, which we regard as a requirement for a proper implementation of such strategy, as most do not allow us to store the data outside of the IDS. We are currently investigating legally in how far storing the data, possibly encrypted, at a co-location would

violate license terms. Eventually, we will have to negotiate license upgrades to explicitly allow storing the data off-site for archival and backup purposes. Moreover the institute is involved in digital preservation activities, e.g., in the context of *nestor*[1] and *WissGrid*[2].

## 5. Using DeReKo

As DeReKo is not available for download, before even mentioning re-usability and sustainability it is, of course, most important to offer a software to access it that fulfils the needs of the target communities. The current corpus search analysis and management system COSMAS II, with currently about 18,500 registered users, offers a broad range of features. E.g., it allows for the composition of virtual corpora, 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.

However, COSMAS II was designed in 1993 for a target corpus size of 300 million words and the growth of DeReKo is pushing it towards its limits. Adding more annotation layers to DeReKo will make the situation even worse.

For that reason we currently prepare a new mid-scale project to create a new corpus analysis system. The new system will have to face opportunities and challenges coming from the emerging distributed e-infrastructures as well as, of course, scientific requirements. To mention but a few:

- it must be suitable for performing methodologically sound empirical linguistic research
- observed data and interpretations need to be separable
- more data is better data: it must allow for large amounts of textual data and annotations (target values are 30 billion words with 20 annotation layers)
- the query mechanism shall allow for multi-layer queries
- query, analysis and metadata function should be connectable to e-infrastructures
- virtual corpora should be definable on metadata and text-internal properties
- users should be able to work on previous states of the data
- users should be able to persistently register virtual corpora (/collections)
- users should be able to add cumulative annotations
- users should be able to run own programs on the data
- the system must guarantee that no license terms are violated

In direct comparison to mere information retrieval systems or web search engines, which also have to deal with amounts of data in a petabyte range, a corpus analysis system for scientific linguistic research has to meet some additional requirements, as for example (see also Kilgarriff, 2007):

- results must be exact and reproducible
- function words cannot be ignored
- indexing has to deal with very unfavourable key distributions
- data structures are more complex: multiple layers and relations on and among annotations have to be represented
- query language needs to be more powerful
- the order of the presentation of search hits has to be controllable, in particular random samples of hits are required

With these additional requirements, at least some commonly used technical tricks and shortcuts for handling large-scale text databases will not be applicable.

## 6. Conclusion

Working on building up corpora since 1964, the IDS has gathered a lot of experience in handling language resources in a sustainable fashion. Despite all difficulties with copyright and licensing, the IDS was and is able to create a large language data resource, which allows for a more empirical approach towards linguistics. The key requirement of sustainability of DeReKo is a continuous maintenance of both the static and the dynamic language resource components and its usefulness for and its usability by its target community, i.e. empirical linguists working on German. To ensure this also for the future, the IDS will start to develop a new corpus management and analysis software. Moreover, the IDS is involved in different infrastructure activities towards sustainability and accessibility of language resources, e.g. in the nestor initiative, WissGrid, TextGrid, and CLARIN.

## 7. References


Belica, C., Kupietz, M., Lüngen, H., Witt, A. (to appear in 2010). The morphosyntactic annotation of DeReKo: Interpretation, opportunities and pitfalls. In: Konopka, M., Kubczak, J., Mair, C., Šticha, F., Wassner, U. (eds), Selected contributions from the conference Grammar and Corpora 2009, Tübingen. Gunter Narr Verlag.

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

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

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based encoding standard for linguistic corpora. In: Proceedings of the Second International Language


[1] nestor – German competence network for digital preservation: http://www.langzeitarchivierung.de/eng/

[2] WissGrid – Grid for Science: http://www.wissgrid.de/index_en.html

Resources and Evaluation Conference (LREC'00), Paris. European Language Resources Association (ELRA).

IDS (1991–2008): COSMAS I/II (Corpus search, Managment and Analysis System). Institut für Deutsche Sprache. Mannheim. http.//www.ids-mannheim.de/cosmas2/

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

IDS (2010). Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010-I (Released 02.03.2010). Institut für Deutsche Sprache. Mannheim. 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.

Kilgarriff, A. (2007). Googleology is Bad Science. In: Computational Linguistics 33 (1). S. 147–151.

Kupietz, M., Keibel, H. (2009). The Mannheim Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research. In: Working papers in corpus-based linguistics and language education. Tokyo University of Foreign Studies (TUFS)

Kupietz, M., Witt, A., Belica, C., Keibel, H. (2010). The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research. In: LREC 2010 Main Conference Proceedings. Malta

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

Sun, S., Lannom, L., Boesch, B. (2003). Handle System Overview. Number 3650 in Request for Comments. IETF, 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/.

Witt, A., Rehm, G., Hinrichs, E., Lehmberg, T., Stegmann, J. (2009). SusTEInability of linguistic resources through feature structures. In: Literary & linguistic computing : LLC ; journal of the Association for Literary and Linguistic Computing, 24 (2009) 3, pages 363–372