Heterogeneity and Standardization in Data, Use, and Annotation: a Diachronic Corpus of German*

Anke Lüdeling
Humboldt-Universität zu Berlin

This paper describes the standardization problems that come up in a diachronic corpus: it has to cope with differing standards with regard to diplomaticity, annotation, and header information. Such highly heterogeneous texts must be standardized to allow for comparative research without (too much) loss of information.

1 Introduction

Most of the corpora in linguistics are fairly large corpora of modern languages (or language stages) that are characterized by a high degree of standardization in three areas: (a) the input data, (b) the annotation, and (c) the intended use.

Input Data In modern languages/language stages orthography is standardized, and texts of the same text type adhere to certain conventions or rules, which makes these texts similar to each other. Most of the tools and mechanisms for collection, pre-processing and evaluation of corpora—symbolic or quantitative—exploit such regularities.

Annotation Most corpora are also standardized with respect to the annotation—be it header information, positional annotation, or structural annotation. Standardization efforts like TEI, OLAC, or IMDI[1] concentrate on these layers.

---

* I want to thank Stefanie Dipper, Lukas Faulstich, Michael Götze, Ulf Leser, Thorwald Poschenrieder, the DDD project members, and the audience of the Potsdam workshop for valuable comments.

Interdisciplinary Studies on Information Structure 02 (2005): 43-54
Dipper, S., M. Götze and M. Stede (eds.):
Heterogeneity in Focus: Creating and Using Linguistic Databases

Intended Use Corpora are usually collected with a given goal in mind and are tailored to fit that goal. This is true even of seemingly 'multi-purpose' corpora like the British National Corpus or the American National Corpus, which are collected specifically for synchronic linguistic research.

On the other hand, we have many less 'well-behaved' corpora: corpora of less studied languages with sometimes little orthographic standardization, or corpus collections that encompass different languages and annotation needs that do not easily conform to corpus-linguistic standards.

The goal of this paper is the description of the problems arising in standardizing the highly variable and heterogeneous data for a diachronic corpus of German. I point to solution strategies and approaches for the representation of the data.

In the following section, I describe the characteristics of our corpus DEUTSCHDIACHRONDIGITAL, a diachronic corpus of German. In Sec. 3, I argue that in addition to a maximally flexible corpus architecture and data model we need a 'corpus model' that ensures standardization and homogeneity wherever possible.

2 DeutschDiachronDigital

The project DEUTSCHDIACHRONDIGITAL[2] (henceforth DDD) aims at the collection, annotation, and presentation of a diachronic corpus of German, covering

[1] The URLs and references for all corpora, tools and other resources mentioned in this paper are given in Section 5.2.

[2] DeutschDiachronDigital is a Germany-wide interdisciplinary project with 12 partners (universities and research institutions). It is still in its beginning phase, with the final funding decision pending. The architecture was developed in a preparatory project funded by the Senatsverwaltung für Wissenschaft, Forschung und Kultur, Berlin. For more information refer to http://www.deutschdiachrondigital.de/.

the language stages Old High German (OHG), Old Saxon (Old Low German; OS), Middle High German (MHG), Early Modern German (EMG), Middle Low German (MLG), and Modern German (MG). DDD thus contains texts from about 800 to about 1900 AD, the focus being on the older texts. The corpus is designed in such a way that it can be used by scholars from (historical) linguistics, (historical) philology, lexicography, history etc. This means that it must be possible to annotate and search for very different kinds of information. Possible research questions include: (i) how did the meaning or form of a specific word change, (ii) how did a given syntactic construction change, (iii) what case do the arguments of a given verb feature, (iv) how did a given genre (let's say, the novel) evolve, (v) what did a given author say about some philosophical concept, (vi) how do letters written by women in the 17th century differ from letters written by men?

At the moment there are quite a number of digitized texts and corpora of older language stages of German, mostly collected in small individual projects with differing standards with regard to diplomaticity, annotation, and header information. Because of these differences, it is at present not possible to conduct qualitative or quantitative research across more than one language stage. A further obstacle is that coverage of the languages stages is very different (for MHG, for example, there are relatively large balanced corpora while for EMG there are almost no electronic resources available, see the survey of Kroyman et al. 2004 for details).

In the following section I describe different kinds of variation within the data and present some ideas of how DDD will cope with them.

3 Standardization Problems and Methods of Resolution

There are in principle two different strategies to deal with the diversity of the data:

(1) normalization or categorization (that is loss of information)

(2) the preservation of different readings[3] (either by explicit coding or by underspecification).

The DDD project uses both strategies, for different types of problems.

3.1 Corpus Architecture

The problem of combining different kinds of texts and annotations in one corpus or database has been tackled in a number of projects (see EXMARaLDA, TUSNELDA, ANNIS etc.; for an overview see Dipper et al. 2004b). Many of them achieve high flexibility by following stand-off models (first developed in multi-modal corpora, see for example Carletta et al. 2003), where the annotation of a text is independent of the text itself and therefore even conflicting hierarchies in different levels of annotation can be accommodated. In some of these tools, the individual texts in the corpus may have differing annotation levels that are in principle totally independent of each other. The flip side of this flexibility is often lack of standardization so that these corpus collections are simply that: collections of texts with no common properties.[4]

---

[3] I use the term ‘readings’ to refer to differences in form or meaning, not just to semantic differences.

[4] This is, of course, due to the situations in which these environments are developed—typically large research groups (Sonderforschungsbereiche) which work on one specific research question in many different languages and approaches; for them it is not a requirement that these resources be directly comparable.

Another problem in conjunction with systems like EXMARaLDA, TUSNELDA, or ANNIS is that they are solutions to very specific problems and cannot easily be transferred to other situations.

For the DDD corpus architecture we use a multi-layer stand-off corpus design with a diplomatic text version as the timeline. Thus, our architecture is as flexible as EXMARaLDA etc.: new texts and new annotation layers can be added at any time. The corpus is stored in a central relational database, import and export to different XML formats is provided via web-clients. The data model is based on an ODAG (ordered directed acyclic graphs) model (see Carletta et al. 2003). Details of the DDD corpus architecture are described in Dipper et al. (2004a) and Faulstich, Leser & Lüdeling (2005). In the remainder of this paper I want to focus on standardization.

3.2 Input Data: Non-Standardized & Multilingual Data

DDD is a historical and diachronic corpus. Historical corpora, even if they consist of texts of one period only, have to deal with non-standardized texts. Not only are there no or little orthographic conventions (depending on the age of the text), there are also many special characters, abbreviations, etc. that are particular to one text. For some historical periods of German (e.g. MHG) it has long been customary to normalize in editions and textbooks. Normalization has a number of advantages: it facilitates readability, and eases access for lexicographical purposes, etc. However, it 'throws away' information about spelling differences and paleographic specifics. Most existing historical corpora designed for linguistic purposes normalize to a certain extent (e.g. the HELSINKI CORPUS) or digitize from already normalized editions of the text instead of original manuscripts or early prints (Rissanen et al. 1993). An exception is the MENOTA PROJECT, which aims at high diplomaticity and is therefore well-suited for paleographical research (however, because corpus composition is not standardized it is less suited for linguistic or lexicographical questions).

Instead of opting for one or the other variant, DDD's multi-layer architecture refers to a highly diplomatic version of the text as a sort of time line and aligns a semi-diplomatic version.

Often there is more than one witness (manuscript, copy) for a given text. DDD digitizes from originals (manuscripts or early prints) rather than critical editions and the focus is on representativity. Hence, it is often necessary to pick one manuscript out of several candidates. In some special cases, however, we will digitize two or more witnesses of the same text (for example, two manuscripts or a manuscript and a critical edition). In these cases we treat each text as a separate text, which comes with its own annotation layers. The witnesses can then be aligned.

Besides being a historic corpus, DDD is a diachronic corpus. A diachronic corpus can be seen as a multilingual corpus (with some parallel portions due to texts that exist across different language stages, such as biblical texts). In addition to standardization on one level, a multilingual corpus has to deal with standardization across different levels, which causes standardization problems in particular at the level of annotation.

3.3 Annotation

Most texts will be annotated with basic structural information, part of speech, lemma, and inflectional morphology. The most relevant standardization problems arise from the fact that tags in any tag set will change their denotation over time. For example, the properties of what would be classified as an 'adverb' are not stable from OHG to MG. This will be dealt with in two ways: the tag sets will be built up hierarchically so that information can be left underspecified if necessary. It will also be possible to explicitly code alternatives.

Standardizing the annotation of lemmas causes particular problems. Ideally, lemmas should be standardized within each language level. Because of the orthographic variance (there are at least 17 spellings of the lemma *und* 'and' in MHG: *undi, unnti, vnnti, vnte, ...*), this is difficult. For some language stages (especially for MHG), there is a well-established normalizing tradition that can be adopted. For other stages, standards have to be developed; cf. the situation in OHG: it is customary to base the normalization on the lemmas as they occur in the largest available OHG text, 'Tatian'. However, this means that many lemmas (namely all those that do not occur in Tatian's text) have to be made up 'in the way Tatian would have written them'. Poschenrieder (2004) suggests a different way of normalizing by representing different but related sounds by abstract 'hyper letters'.

There are no conventions for lemma correspondence that hold across all language stages. Lexical change occurs on the semantic, morphological, and stratic level (or combinations thereof, see Gévaudan 2002 and Gévaudan & Wiebel 2004 for a discussion and a modelling proposal). Therefore, it is not a trivial task to decide which elements (i.e., normalized lemmas) correspond to each other across language stages.

3.4 Intended Use: Corpus Composition

As stated above, DDD will be used by scholars from different fields such as (historical) linguists, (historical) philologists, lexicographers, historians etc. as well as by interested laypeople. In this way it already differs from most available corpora. The existing historical and diachronic corpora are usually compiled either for linguistic purposes or for special philological, lexicographic, or historical purposes (for an overview over historical and diachronic corpora see Kroymann et al. 2004). This is reflected by the *corpus composition*: historical corpora for linguistics or lexicography, such as the HELSINKI CORPUS or the corpora for the MITTELHOCHDEUTSCHES WÖRTERBUCH, are in some way 'representative', e.g., they cover language stages, authors, genres, etc. in given proportions (Klein 1991, Biber 1993). Corpora for philological purposes, on the other hand, are often much more specialized—they cover the work of one author only, or sometimes even only one work (the CANTERBURY TALES PROJECT), or one genre (LANCASTER NEWSBOOK CORPUS), for an overview see Burch et al. (2003).

To make diachronic research possible, a diachronic corpus must be 'representative' with respect to time, dialect, text type, etc. This is difficult to achieve in diachronic corpora because categories like 'text type' or 'dialect' are not stable across time. Some genres or text types[5] only develop during the sampling time (like the category 'novel') and others appear and then disappear again (like minne songs). Even if the categories were stable and one could draw a matrix of different parameters, it would not be possible to fill all the cells in such a matrix because there is simply not enough material (this is, of course, especially true for the early stages).

The DDD project deals with these problems in two ways: a very detailed common set of parameters (time, genre, dialect, information about the author, register, purpose of the text etc.) is chosen, which is represented in hierarchies so that it is possible to always use the most specific category. The categories form a matrix, and corpus selection aims at filling as many of the matrix cells as possible (many will remain empty). If there are several texts that fit a cell, the most well-known is chosen.

For the early language stages (OHG, OS), all available texts are included in the corpus. From the time of MHG/MLG on, there are too many texts avail-

[5] The problem of defining a 'genre' or 'text type' is ignored here. I assume that there is some kind of available definition.

able so that there has to be a selection. For even later language stages (older Modern German), the matrix would become too complex since many more genres develop. Therefore DDD concentrates on three genres (letters, newspaper text, novels) and leaves other cells empty (they can be filled later).

The categories and their values are represented in a TEI-conform header so that it is possible to construct sub-corpora for different purposes.

4 Conclusion

In order to make a corpus out of very different texts from different language stages, a maximally flexible corpus architecture is necessary as well as standardization in many ways. In a diachronic corpus, it must be possible to make use of as much information as possible from every text (because there are so few texts to begin with).

DDD deals with this situation by using a multi-layer architecture where text versions of different degrees of diplomaticity are aligned (see Lüdeling, Poschenrieder & Faulstich 2005). The tag sets are hierarchically ordered so that information can be left underspecified where necessary. In addition, different hypotheses can be coded explicitly. This architecture is able to cope with many of the challenges arising from the specific needs of diachronic data.

5 References

5.1 Bibliography

Biber, Douglas (1993) Representativeness in Corpus Design. In: *Literary and Linguistic Computing* 8, 243–257.

Burch, Thomas; Johannes Fournier; Kurt Gärtner & Andrea Rapp (eds.) (2003): *Standards und Methoden der Volltextdigitalisierung. Beiträge des Inter-*

nationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2001. Akademie der Wissenschaften und der Literatur, Mainz.

Carletta, Jean; Jonathan Kilgour; Timothy O'Donnell; Stefan Evert & Holger Voormann (2003) The NITE object model library for handling structured linguistic annotation on multimodal data sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web*, Budapest.

Dipper, Stefanie; Lukas Faulstich; Ulf Leser & Anke Lüdeling (2004a) Challenges in Modelling a Richly Annotated Diachronic corpus of German. In *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora,* , pp. 21–29, Lisbon.

Dipper, Stefanie; Michael Götze & Manfred Stede (2004b) Simple Annotation Tools for Complex Annotation Tasks: an Evaluation. In *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora,* pp. 54–62, Lisbon.

Faulstich, Lukas; Leser, Ulf & Lüdeling, Anke (2005) *Storing and Querying Historical Texts in a Database*. Technical Report 176 of the Institut für Informatik. Humboldt-Universität zu Berlin. Available online at http://www.deutschdiachrondigital.de/publikationen/index.php

Gévaudan, Paul (2002) *Klassifikation des lexikalischen Wandels. Semantische, morphologische und stratische Filiation*, Dissertation, Universität Tübingen.

Gévaudan, Paul & Dirk Wiebel (2004) Dynamic lexicographic data modelling. A diachronic dictionary development report. In *Proceedings of the LREC Conference,* Lisbon.

Klein, Thomas (1991) Zur Frage der Korpusbildung und zur computergestützten grammatischen Auswertung mittelhochdeutscher Qellen. In: Wegera, Klaus-Peter (ed.) *Mittelhochdeutsche Grammatik als Aufgabe*. Zeitschrift für deutsche Philologie 110 (Sonderheft) 1991, 3–23.

Kroymann, Emil; Sebastian Thiebes; Anke Lüdeling & Ulf Leser (2004) *Eine vergleichende Analyse von historischen und diachronen digitalen Korpora*. To appear as Technical Report of the Institut für Informatik, Humboldt-Universität zu Berlin. Available online at http://www.deutschdiachrondigital.de/publikationen/index.php.

Lüdeling, Anke; Thorwald Poschenrieder & Lukas Faulstich (2005) DeutschDiachronDigital. Ein diachrones Korpus des Deutschen. In *Jahrbuch für Computerphilologie 2004.* Available online at http://www.deutschdiachrondigital.de/publikationen/index.php

Poschenrieder, Thorwald (2004) *Digitalisierung altgermanischer Texte*. Paper given at the Conference of the Society for Indo-European Studies, Krakau, http://www.deutschdiachrondigital.de/publikationen/index.php

Rissanen, Matti; Merja Kytö & Minna Pallander (1993) *Early English in the Computer Age: Explorations through the Helsinki Corpus,* Mouton de Gruyter, Berlin.

5.2 Corpora and Tools[6]

- ANNIS (a Linguistic Database for Exploring Information Structure): http://www.sfb632.uni-potsdam.de/annis/
- The CANTERBURY TALES PROJECT: http://www.cta.dmu.ac.uk/projects/ctp/
- EXMARaLDA (Extensible Markup Language for Discourse Annotation): http://www.rrz.uni-hamburg.de/exmaralda/
- The Diachronic Part of the HELSINKI CORPUS, delivered by ICAME, user manual: http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM
- IMDI (Isle Metadata Initiative): http://www.mpi.nl/IMDI/
- The LANCASTER NEWSBOOK CORPUS:

[6] The URLs given below were valid URLs on March 11, 2005.

http://bowland-files.lancs.ac.uk/newsbooks/project.htm

- MITTELHOCHDEUTSCHES WÖRTERBUCH: http://gaer27.uni-trier.de/MWV-online/MWV-online.html
- The MENOTA PROJECT: http://gandalf.aksis.uib.no/menota/
- NITE http://nite.nis.sdu.dk/aboutNite/
- OLAC (Open Language Archives Community): http://www.language-archives.org/
- TEI (Text Encoding Initiative): http://www.tei-c.org/
- TUSNELDA (Tübinger Sammlung nutzbarer empirischer, linguistischer Datenstrukturen): http://www.sfb441.uni-tuebingen.de/tusnelda.html

Anke Lüdeling
Korpuslinguistik
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
anke.luedeling@rz.hu-berlin.de
http://www.linguistik.hu-berlin.de/korpuslinguistik/