Annotating a Multi-Genre Corpus of Early Modern German

Paul Bennett, Martin Durrell, Silke Scheible and Richard J. Whitt
School of Languages, Linguistics, and Cultures
University of Manchester
Paul.Bennett@manchester.ac.uk
Martin.Durrell@manchester.ac.uk
Silke.Scheible@manchester.ac.uk
Richard.Whitt@manchester.ac.uk

Abstract

This study addresses the challenges in annotating a spatialised multi-genre corpus of Early Modern German with linguistic information, and describes how the data can be used to carry out a systematic evaluation of state-of-the-art corpus annotation tools on historical data, with the goal of creating a historical text processing pipeline.

The investigation is part of an ongoing project funded jointly by ESRC and AHRC whose goal is to develop a representative corpus of Early Modern German from 1650 to 1800. This period is particularly relevant to standardisation and the emergence of German as a literary language, with the gradual elimination of regional variation in writing. In order to provide a broad picture of German during this period, the corpus includes a total of nine different genres, and is ‘spatialised’ both temporally and topographically, i.e. subdivided into three 50-year periods and the five major dialectal regions of the German Empire. Due to the lexical, morphological, syntactic, and graphemic peculiarities characteristic of this particular stage of written German, and the additional variation introduced by the three variables of genre, region, and time, automatic annotation of the texts poses a major challenge.

1. Introduction

The aim of the GerManC project is to compile a representative corpus of Early Modern German covering the years 1650-1800. The corpus is modelled on the ARCHER corpus for English, which aims to be a representative corpus of historical English registers and consists of samples of 2000 words of continuous texts for a number of genres/registers.

The GerManC corpus was born out of the need for a resource to facilitate comparative studies of the development and standardisation of English and German in the 17th and 18th centuries. PhD students at the University of Manchester who were working on aspects of the history of German and comparative historical studies of English and German were obliged to assemble their German data manually, with concomitant problems in terms of time, representativeness and comparability (Storjohann, 2003; Auer, 2009). The early modern period (1650-1800) is a crucial one in the emergence of German as a literary language, with the development of codification and the acceptance of a standard written form of German throughout the German-speaking lands (Blackall, 1978). The developments over this period offer a fruitful basis for comparison with developments in English, notably in understanding the processes which contributed to linguistic standardisation. In syntax, for example, developments took place at this time which are now considered to constitute the most significant differences between English and German, such as the fixing of clause-final position for certain verb forms in German. Aside from this linguistic interest, such a corpus is potentially a valuable interdisciplinary resource, providing material for literary, cultural and social historians.

The current project, which started in 2008, was preceded by a one-year pilot study (also funded by the ESRC), whose purpose was to test the corpus design and aims, and involved compilation of one genre (newspapers). We are now in the process of adding further genres and linguistic annotation, which is the topic of this paper.
2. Corpus design

This section describes the GerManC corpus in more detail, as its unusual structure represents a significant challenge for automatic annotation. What makes it so special is that it aims to be representative on three different levels. First of all, it is desirable for it to be representative of language usage in this period. In order to achieve this, the corpus includes a range of registers or text types and, as far as possible, each register is represented by a sample of equal size. This means that the corpus does not consist of complete texts (which could mean that one text type, for example long novels, would be overrepresented), but of relatively short samples. The sample size of the Brown and ARCHER corpora, with extracts of some 2000 words (Meyer, 2002: 40-42), has proved its viability over time, and we decided to follow this model. GerManC thus includes nine different genres, which are modelled on the ones used in ARCHER: four orally-oriented genres (dramas, newspapers, letters, and sermons), and five print-oriented ones (journals, narrative prose, scholarly writing in the humanities, scientific texts, and legal texts).

Secondly, in order to enable historical developments to be traced, the period is divided into fifty year sections (in this case 1650-1700, 1701-1750 and 1751-1800), and an equal number of texts from each register selected for each of these periods. This periodisation follows the model established for the Bonn corpus (Hoffmann and Wetter, 1987), which formed the basis for the grammar of Early New High German by Besch, Moser & Stopp (1970), and this proved to be adequate to capture chronological variation at this time. The combination of historical and text-type coverage should enable research on the evolution of style in different genres, along the lines of previous work for English (Atkinson, 1992; Biber and Finegan, 1989).

Finally, the sample texts also aim to be representative with respect to region. This dimension has not been seen as essential for English corpora. ARCHER, for instance, only considers the two varieties British English and American English, but no further regional variation among these areas. The reason why different speech areas are taken into account in GerManC is that regional variation remained significant much longer in the development of standard German than it did in English (Durrell, 1999). However, this variation diminished over the period in question as the standard originating in the Central German area was gradually adopted in the South. Enabling this development to be traced systematically is one of the crucial desiderata for this corpus. Figure 1 shows the five broad regional areas that are included in the corpus, i.e. North German, West Central German, East Central German, West Upper German (including Switzerland), and East Upper German (including Austria).

![Figure 1: Dialect regions in GerManC](image)

Altogether, per genre, period, and region, around three extracts of at least 2000 words are selected, yielding a corpus size of around 900,000 words altogether. This includes the
newspaper data compiled in the pilot project. The structure of the corpus is summarised in Table 1.

<table>
<thead>
<tr>
<th>Periods</th>
<th>1650–1700</th>
<th>1701–1750</th>
<th>1751–1800</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regions</td>
<td>North</td>
<td>West Central</td>
<td>East Central</td>
</tr>
<tr>
<td>Genres</td>
<td>Orally-oriented</td>
<td>Print-oriented</td>
<td></td>
</tr>
</tbody>
</table>

*Table 1: Structure of the Corpus*

3. Digitisation and structural annotation

After texts have been selected, they are transcribed manually. A manual approach to digitisation was chosen because the majority of texts is printed in black letter fonts of variable sizes (*Fraktur*), as illustrated in Figure 2. The pilot study showed that scanning *Fraktur* with OCR technology is impractical and prone to error, especially when considering that text samples are taken from a variety of genres, and printed in different locations. Further problems are the arbitrary variation in font size, the denseness of the print on the page in many texts, and frequent variation between black letter and Roman fonts, even within words (as exemplified by the newspaper text at Figure 4). The most reliable method for the digitisation of such older texts is by means of double-keying, i.e. each text is keyed in by two individuals and the results compared electronically to eliminate errors. This technique was adopted for the pilot project and found to be wholly satisfactory.

![Figure 2: Drama excerpt](image)

The raw input texts are then annotated according to the guidelines of the Text Encoding Initiative (TEI). The Text Encoding Initiative has published a set of XML-based encoding...
conventions recommended for meta-textual markup in corpus projects around the world and across different computer systems (http://www.tei-c.org). The principal aim of this is to minimise inconsistencies across projects and to maximise mutual usability and data interchange. Historical text annotation in particular needs to be very detailed with regard to document structure, glossing, damaged or illegible passages, foreign language material and special characters such as diacritics and ligatures. For the purpose of annotating these issues in our texts we decided to use the TEI P5 Lite tagset. Mark-up is carried out using the OXygen XML editor, which provides special support for TEI annotation.

Figure 3 shows a drama excerpt which has been structurally annotated, showing headers, stage directions, speakers (including a “who” attribute for co-reference), as well as lines. One of the insights gained from the pilot project (in which newspapers were annotated according to TEI) was that this annotation should not be too detailed. The TEI tagset (and even the “Lite” version) is so rich that annotators were tempted to spend too much time on structural annotation, i.e. annotation of non-linguistic features which are not a priority for this project. As GerManC is a corpus aimed at linguists in particular, the focus of the main project is directed towards linguistic annotation.

4. Linguistic annotation

In order to facilitate a thorough linguistic investigation of the data, we plan to add the following types of annotation:

- Word tokens
- Sentence boundaries
- Lemmas
- POS tags
- Morphological tags

The central question with respect to these plans is to what degree annotation can be automated. That is, are there any tools available that can adequately handle the data in our corpus?

Before describing the challenges in annotating our corpus, one issue that needs to be addressed in advance concerns the choice of annotation format, as there is still no agreed norm for linguistic annotation. Ideally, linguistic annotation should simply be integrated into the TEI markup described in the previous section. However, one commonly known problem with TEI is that its inline XML format does not lend itself very well to more complex annotations. For example, it is not always possible to add sentence tags to drama excerpts that involve line or line group markup, as can be seen in Figure 3. Here, the sentence “wiewol es
One solution to this is to keep structural TEI annotations and linguistic annotations separate. This, however, would result in two separate versions of the corpus, which is not desirable. The second (and our preferred) alternative is to work with a stand-off annotation format, such as the ones employed by GATE (http://gate.ac.uk/) or UIMA (http://incubator.apache.org/uima/). In addition to allowing several layers of annotation, stand-off formats also allows annotations to be easily imported into a relational database such as MySQL (http://www.mysql.com/), which we plan to make accessible via an online interface at the end of our project.

In general, annotating historical texts with linguistic information is not straightforward. Our data is problematic on a number of levels. Firstly, there is variation on the three extra-linguistic levels of genre, region and time. Secondly, the period of Early Modern German is characterised by a considerable amount of variation on all levels of language, including orthography, lexis, morphology, and syntax.

First of all, there are some basic orthographic challenges, which are bound to have an effect on the performance of even the lowest-level processing tools. For example, punctuation is far from standardised and may vary not only over different genres but also over time and even within a single text. For example, the virgule symbol "/" survived longer in German than in English, and was used to separate textual segments of varying length and grammatical status (compare the drama excerpt in Figure 3 where the virgula is used mark rhythm in verse with the newspaper text in Figure 4, where it is primarily used to mark the end of clauses). In newspaper texts it disappeared rather quickly, being replaced mainly by the comma: while it is still common in a text from 1744, it is found only once in the third period, in a text from 1793, where it occurs in a header to separate place and date, which means that it is not used as punctuation in running text. In addition, the virgule symbol is often used in place of both comma and full-stop, which makes it difficult to identify sentence boundaries. This is particularly relevant for dramas and academic texts, where virgules are used alongside commas and periods, and it is not always apparent which punctuation mark is supposed to be serving which function.

![Figure 4: Newspaper text (Aussfuehrliche Relation, Breslau, 1683)](image-url)
Two further problems result from printing rather than punctuation choices. Firstly, word boundaries are at times hard to determine as printers often vary in the amount of whitespace they leave between two words. For instance, sometimes they attempt to squeeze in an extra word at the end of a line, and as a result it is not straightforward to determine if one or two words were intended. Secondly, sometimes whole passages or words are hard to read because of bad print quality. Such cases are marked with "<unclear>" tags in our data.

On a lexical level, automatic tools have to cope with a wealth of unknown words, including obsolete words, loan words and nonce-formations (i.e. words which may only occur once ever). In addition, texts are characterised by large amounts of regionalisms, such as the German word for “goat”, Ziege, which in the south of Germany is Geiß. One further difficult issue is the fact that the data show great variation in spelling. While some cases are fairly regular and could be captured by simple substitution rules (e.g. single vs. double f in German “run”, lauff(en), others are not so easy to predict, as for example the German word for “terrible”, which occurs in our data both in its modern form schrecklich, but also as the variant schrölick. Spelling variation can be particularly extensive for named entities: in the newspaper data alone the city of Madrid occurs with seven different spellings. Finally, there is considerable variation in the application of apocope, i.e. whether a final -e is written at the end of a word or not, as for example in German “today”, heute vs. heut. In general, texts from the South tend to omit -e more frequently than those from more northerly regions, but this is by no means consistent. All of these issues (summarised in Table 2) are difficult to deal with for automatic tools (especially for a lemmatiser and POS-tagger).

<table>
<thead>
<tr>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown words</td>
</tr>
<tr>
<td>Regionalisms</td>
</tr>
<tr>
<td>Spelling variation</td>
</tr>
<tr>
<td>Apocope</td>
</tr>
</tbody>
</table>

Table 2: Lexical challenges

On a morphological level, variation occurs in plural formation, as for example in German “day”, Tag, which occurs with three different plural forms: Tage, Tag und Täg. A further example is German “harbour/port”, which occurs with plural forms Hafen, Hafens, and Häfen. In both examples the first option has become standardised. Further morphological variation can be found for example in the weak adjective inflection in nominative/accusative plural position, where there is variation between forms with final -en, and forms without: ‘die guten Kinder’ vs. ‘die gute Kinder’ (see also Voeste (1999)). The form with -en is now standard, and -e survives only in south-western dialects. In our pilot corpus, -e is dominant in the first period, except in East Central German; in the second period, -en is dominant in both North and East Central German. By 1800, -en is dominant everywhere, except in the south west. The former is shown to be in the process of standardisation during this period.

<table>
<thead>
<tr>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plurals</td>
</tr>
<tr>
<td>Hafen (sg.) – Hafen / Hafens (pl.)</td>
</tr>
<tr>
<td>Weak adjective inflection</td>
</tr>
</tbody>
</table>

Table 3: Morphological challenges

On the level of syntax, one major issue is sentence length, which varies greatly among the different genres. For example, newspapers tend to have very long and complex sentences which often show unusual orders of verb elements. For example, a finite modal verb and a passive infinitive can occur in the subordinate clause with the finite verb first, as in ‘[...] dass es [...] soll gemacht werden’, as opposed to the modern variant ‘[...] dass es [...]
*gemacht werden soll* (“[...] that it should be done”). The first variant is frequent until 1750, and only after this does the second variant become clearly dominant.

<table>
<thead>
<tr>
<th>Sentence length</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order of verb elements</td>
<td>a.) [...] dass es [...] soll gemacht werden</td>
</tr>
<tr>
<td></td>
<td>b.) [...] dass es [...] gemacht werden soll</td>
</tr>
<tr>
<td>Suppression of auxiliaries</td>
<td></td>
</tr>
<tr>
<td>Relative pronoun ‘so’</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Syntactic challenges

Further syntactic challenges are posed by suppressed auxiliaries, i.e. the almost regular omission of the perfect auxiliary haben or sein in subordinate clauses, and the use of ‘so’ as a relative pronoun. Both are illustrated by the following sentence taken from “Altonaischer Mercurius”, a newspaper printed in Altona in 1698:


‘A raiding party from Caransebes defeated an enemy raiding party which had brought corn to Temesvar, killed many and captured six Janissaries with 300 horses.’

*Altonaischer Mercurius, Hamburg Altona, 15/11/1698*

The example shows an instance of “so” used as relative pronoun, and a missing auxiliary “hat” has also been highlighted. The sentence further illustrates a number of other issues described above, such as spelling variation in “Parthey”, which in modern German is spelled “Partei”. There is also a likely case of regional variation: the word “Früchte” means “fruit” in modern German, but in this context it is likely to mean “corn”, a regional usage particularly frequent in the South-East where the report originated. Finally, there is also a morphological variant of a verb in the sentence, “erleget”, which in modern German would be “erleget”. This exemplifies syncope - the loss of a medial vowel, which is again inconsistent at this period.

The examples in the previous paragraphs aim to illustrate the complexity of the data that automatic tools will have to deal with. The main problem is that only few specialised tools are available for processing historical corpora and other non-standard varieties of language. With respect to current tools, on the other hand, the question is to what extent they are suitable for our kind of data, as they are generally tuned to perform well on the type of data they have been developed on, but are likely to perform much worse on other types of data. This is a general problem that linguists working with non-standard language have to deal with.

6. Goals

As our corpus displays such a wealth of variation, it lends itself as an ideal test bed for evaluating current corpus annotation tools. This will be of particular interest to historical corpus linguists who are faced with the difficult decision of which tools are most suitable for processing their data, and are likely to require the least manual correction.

The second goal then utilises the findings of this evaluation study to improve the performance of existing tools, with the goal of creating a historical text processing pipeline, which will contain the following components:
- Tokeniser
- Sentence boundary detector
- Lemmatiser
- POS tagger
- Morphological analyser

It will also have options to specify the type of input data present. That is, the components of the pipeline are tuned to deal with genre variation, and possibly also temporal and spatial variation in Early Modern German.

Finally, our planned contributions involve the following. First of all, it is of great importance to define clear annotation guidelines for all proposed linguistic annotation tasks. We have realised that this is already a fairly complex task with respect to tokenisation. Further issues will involve the question whether modern tagsets such as the STTS part-of-speech tagset for modern German are actually appropriate for historical data (Schiller et al., 1999). On the basis of these findings, we are also preparing a substantial gold-standard data set which is manually annotated by two team members and can be used as test bed for current tools. We are hoping that it will be big enough to also allow for experimenting with retraining and testing tools such as POS taggers.

We are planning to use the gold-standard data to test current tools and evaluate their suitability for historical data. Various aspects will be considered, such as their performance on particular genres, periods, or regions as well as their overall performance and robustness. However, our main focus will be on creating a historical text processing pipeline by improving and adding current tools and by implementing new ones. We have started work on this by implementing a novel tokeniser and sentence boundary detector which are tuned to deal with the variety of data in our corpus, and which also have a number of output options suitable for further processing. We also plan to experiment with a novel spelling variant detector for Early Modern German which can identify the modern spelling of earlier non-standard variants in context, similar to Rayson et al.’s variant detector tool (VARD) for English (2005), and complementing the work of Ernst-Gerlach and Fuhr (2006) and Pilz et al. (2006) on historic search term variant generation in German.

References


