Annotating a historical corpus of German: A case study

Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt

University of Manchester
Oxford Road, Manchester, M13 9PL

Abstract

We report on our experiences in annotating a historical corpus of German with structural and linguistic information, providing an example of the needs and challenges encountered by smaller humanities-based corpus projects. Our approach attempts to follow current standardisation efforts to allow for future comparative studies between projects and the potential extension of our annotation scheme. Structural information is encoded according to TEI (P5) guidelines, and the corpus is further being annotated with linguistic information in terms of word tokens, sentence boundaries, normalised word forms, lemmas, POS tags, and morphological tags. The major problem encountered to date has been how to merge the linguistic mark-up with the TEI-annotated version of the corpus. In the interest of interoperability and comparative studies between corpora we would welcome the development of clearer procedures whereby structural and linguistic annotations might be merged.

1. Introduction

GerManC is an ongoing project based at the University of Manchester and funded jointly by the ESRC and AHRC. Its goal is to develop 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 continuous texts for a number of genres/registers. GerManC includes nine different genres and is subdivided into three 50-year periods and the five major dialectal regions of the then German Empire. Like ARCHER, it consists of sample texts of 2,000 words (yielding 900,000 words altogether), and two-thirds of the digitisation is now complete.

We shall report on our experiences in annotating the GerManC corpus with structural and linguistic information, providing an example of the needs and challenges encountered by smaller humanities-based corpus projects. As we are collaborating with various other historical projects that are currently in progress (for example, addressing other stages of German1, or other languages, Biber et al. (1994)), it is of major importance to choose a standardised annotation format to enable interoperability and comparison. Our approach therefore attempts to follow current standardisation efforts to allow for future comparative studies between projects and the potential extension of our annotation scheme2. Our results will be of particular interest to related projects which still use their own specialised annotation formats.

2. Corpus compilation and design

As GerManC is a historical corpus which will primarily be used in corpus linguistic studies, its design and annotation needs differ significantly from current large-scale corpus compilation projects. First of all, digitised historical data from the Early Modern German period is scarce, which means that the majority of texts included in the corpus have to be digitised first. A manual approach to digitisation was chosen as texts from this period are usually printed in black letter fonts of variable sizes (Fraktur), as illustrated in Figure 1. Initial tests showed that scanning Fraktur with OCR technology is impractical and prone to error, especially because 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. 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 project and found to be wholly satisfactory.

Although representativeness is difficult to achieve, GerManC aims to provide a broad picture of Early Modern German and takes three different levels into account. First of all, 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

---

1http://www.linguistics.ruhr-uni-bochum.de/dipper/projectddd.html
2For example, to include syntactic mark-up.

---

Figure 1: Drama excerpt
words (Meyer, 2002), 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 the corpus includes an equal number of texts from each register for each of these periods. The periodisation follows the model used in the Bonn corpus (Hoffmann and Wetter, 1987), which proved to be adequate to capture chronologically 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.

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. Although this only a relatively small size, the unusual structure of the corpus represents a significant challenge for annotation.

## 3. Structural annotation

Annotation of historical texts 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 this purpose, the raw input texts are annotated according to the guidelines of the Text Encoding Initiative (TEI) during manual transcription. 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. The principal aim of TEI is to minimise inconsistencies across projects and to maximise mutual usability and data interchange. For the purpose of annotating these issues in our texts we use the TEI P5 Lite tagset as it offers a wealth of strategies for encoding structural details, and serves as standard for many humanities-based projects. Transcription and structural annotation are carried out using the OXygen XML editor, which provides special support for inline TEI annotation. Only the most relevant tags are selected from the set to keep the document structure as straightforward as possible. Due to the great variability of our corpus with respect to different genres, regions, and time periods, the full subset of TEI tags used will only be known at the end of the digitisation stage. Figure 2 shows the structural annotation of the above drama excerpt, including headers, stage directions, speakers (including a “who” attribute for co-reference), as well as lines.

![Figure 2: Structural annotation](http://www.oxygenxml.com)

## 4. Linguistic annotation

The corpus is further being annotated with linguistic information in terms of word tokens, sentence boundaries, normalised word forms, lemmas, POS tags, and morphological tags. To reduce manual labour, a semi-automatic approach was chosen whose output is manually corrected. More detail about the annotation procedure can be found in Section 5.

Each annotation type requires careful consideration and adaptation as German orthography was not yet codified in the Early Modern period. Decisions on the level of tokenisation are especially important, as (with the exception of sentence boundaries) all other annotation types are token-based. 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. Clitics and multi-word tokens are particularly difficult issues: lack of standardisation means that clitics can occur in various different forms, some of which are difficult to tokenise (e.g. *wirstu* instead of *wirst du*). Multi-word tokens, on the other hand, represent a problem as the same expression may be sometimes treated as compound (e.g. *obgleich*), but written separately at other times (ob gleich). While our initial tokenisation scheme takes clitics into account, it does not yet deal with the issue of multi-word tokens. This means that whitespace characters act as token boundaries, and multi-word expressions will be identified in a later step.

Annotation of sentence boundaries is also affected by the non-standard nature of the data. Punctuation is not standardised in Early Modern German and varies 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. It 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 full stops, and it is not always apparent which punctuation mark serves which function. The tokenised text is further annotated with normalised word forms, lemmas, POS tags, and morphological tags. While the latter two tasks use standard tagsets (STTS for POS tagging\(^5\), and extended STTS for morphological information), we have defined special guidelines for annotating normalised word forms and lemmas. The normalisation stage aims to address the great amount of spelling variation that occurs in written historical documents, which proves problematic for automated annotation tools. Before final codification words often appear in a variety of spellings, sometimes even within the same paragraph or text. As most current corpus processing tools (such as POS-taggers or lemmatisers) are tuned to perform well on modern language data which follows codified orthographic norms, they are not usually able to account for variable spelling, resulting in lower overall performance (Rayson et al., 2007). Our goal is to develop a tool similar to Baron and Rayson’s variant detector tool (VARD) for English (Baron and Rayson, 2008), and complementing the work of Ernst-Gerlach and Fuhr (2006) and Pilz and Luther (2009) on historic search term variant generation in German, which will help to improve the output of the POS tagger and lemmatiser and will thus reduce manual labour.

In addition to creating a gold-standard annotation of our corpus which can be used to carry out reliable corpus-linguistic studies of Early Modern German, we also plan to make the following contributions:

- Provide detailed annotation guidelines for all proposed annotations
- Test and evaluate current corpus annotation tools on gold standard data
- Identify techniques for improving the performance of current tools
- Create historical text processing pipeline

To allow for future comparative studies between projects, we provide detailed annotation guidelines for both structural and linguistic annotation. Furthermore, as GerManC displays a wealth of variation in terms of different genres, time periods and regions, it lends itself as an ideal test bed for evaluating current corpus annotation tools (POS taggers etc.). This will be of particular interest to future corpus compilation projects 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 a tokeniser, a sentence boundary detector, a lemmatiser, a POS tagger, and a morphological analyser. It will also have options to specify the type of input data present, i.e. the components of the pipeline will be tuned to deal with genre variation, and possibly also temporal and spatial variation in Early Modern German.

5. Annotation procedure and challenges

In order to create the gold-standard annotation of our corpus and achieve the goals outlined in the previous section, our team was faced with a number of challenges. With no historical text processing platform yet available, we had to identify a suitable framework which would satisfy the following requirements:

1. Automate linguistic annotation (for subsequent manual correction)
2. Provide facilities for manual correction (annotation tool)
3. Produce standardised annotation format (suitable for further processing and comparison with other projects)
4. Merge structural (TEI) annotation with linguistic annotation

We identified GATE (Cunningham et al., 2002) as the most suitable framework for the tasks described above. GATE (“General Architecture for Text Engineering”) is open source software “capable of solving almost any text processing problem”\(^6\). To address point 1.), we used GATE’s German Language plugin\(^7\) and the TreeTagger (Schmid, 1994) to obtain annotations in terms of word tokens, sentence boundaries, lemmas, and POS tags. As GATE also offers facilities for manual annotation, we simultaneously use it as an annotation tool, correcting the errors produced by the automated tools to produce a gold standard annotation (point 2.).

The major problem encountered to date has been how to incorporate linguistic information in the TEI-annotated version of the corpus without invalidating the existing XML structure or ending up with two separate versions of the corpus. Structural and linguistic annotations cannot be merged into an inline XML format, as conflicts arise on a number of levels. For example, the inline XML structure of the drama excerpt in Figure 2 would be invalidated if sentence mark-up was added. Here, the sentence “wiewol es nicht vor mich geschehn/ was ich mit ihr geredt” (“although we did not speak at my behest”) stretches across a line boundary, leading to a crossed (and consequently invalid) XML structure. Furthermore, in words perceived as ‘foreign’, the typeface is frequently changed in the middle of a token, with Roman type used for the ‘foreign’ root and black letter (“Gothic”) for the inflectional ending, as for example in the last word of the line in Figure 3 (marked as <hi>repetir</hi>et). Adding word token mark-up would only be possible if the typeface mark-up <hi> was nested within the token mark-up, which creates a conflict with the requirement that token tags should occupy the lowest level in the hierarchy.

5\(^5\)http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html
6\(^6\)http://gate.ac.uk
7\(^7\)http://gate.ac.uk/sale/tao/splitch19.html#x24-45900019.1.2
The TEI guidelines offer some guidance to annotating cases where there is a conflict between the XML hierarchy established by the physical structure of a text (e.g., paragraphs, lines) and its linguistic structure (e.g., sentences, tokens). One suggested solution is to mark boundaries with empty elements, and including information about the start and end points of non-nesting material. This prevents the document from becoming invalid, and furthermore preserves the beginnings and end points for further processing. The TEI guidelines further discuss the use of stand-off format, in which text and mark-up are separated (for example by using XML elements which contain links to other nodes in another XML document).

Although TEI offers some solutions for merging structural and linguistic annotation, no information is provided on how the required annotations could be added automatically. Crucially, most automatic processing tools for German do not yet support TEI and require plain text as input (e.g., Lemnitzer’s Perl tokenizer, TreeTagger). Given that the structural annotation is added first (in our case, by using inline XML tags during first inputting), a framework is required whereby automated linguistic annotation tools can be run on TEI-encoded texts and merge the newly created linguistic annotations with the existing structural mark-up. It seems that to date no such framework is available, and little documentation is available on how structural TEI annotations can be merged with linguistic annotations. We further found that some annotations proposed in the TEI P5 manual are unsuitable for further linguistic processing, as they allow manipulation of the original document by adding information on the text level. For example, TEI’s treatment of abbreviations suggests the use of a “choice” element to record both the abbreviation and its expansion, as illustrated in Figure 4.

![Figure 4: TEI’s choice tag](image)

Another example is the inclusion of descriptive elements, which can be used to provide short textual descriptions of omitted figures or graph. The implications of such additions on the text level are twofold. First of all, the original text flow is interrupted, which represents a problem for further processing tools such as POS taggers (which would treat the added material as part of the original text). Secondly, the fact that TEI allows manipulation of the original text means that an approach where the structural and linguistic annotations of a text are merged at a later stage is not straightforward as the underlying documents (and character offsets) may differ.

Although GATE does not specifically support TEI, it allows XML-encoded text as input, which means that the above-mentioned German Language plugin can be applied to the TEI-annotated version of our corpus, and both annotation layers can be saved within the same document. This can be achieved in two different ways: by using a GATE-particular stand-off format, or by saving the annotated text as inline XML using the ‘Save Preserving Format’ option (which attempts to preserve the original XML mark-up alongside the new annotations). However, from our point of view both formats are problematic: the stand-off architecture is GATE-specific and needs to be transformed into other formats for external processing (e.g. by using XSLT stylesheets). The inline format, on the other hand, has to deal with overlapping elements as the ones described above (crossed line boundaries in the drama corpus and changes in typeface). Ideally, the ‘Save Preserving Format’ option should address such issues by extending the spans of the original structural mark-up to wrap around the newly created linguistic annotations, and adding information about the start and end points of the non-nesting material (as suggested in the TEI manual, see above). Instead, the original mark-up is manipulated to accommodate the new annotations in a way which can lead to inaccuracies on the structural level. For example, the token `<hi>repetir</hi>` is shown in Figure 3 is wrapped as `<hi><w>repetir</w></hi>` by the tokeniser module, incorrectly splitting the token into two parts. To deal with these issues we created scripts which “repair” such cases by using fragmentation techniques similar to the ones described in the TEI manual.

From the point of view of a smaller humanities-based project it would be desirable if text processing platforms such as GATE provided explicit support for texts encoded according to the TEI P5 guidelines, as a great deal of time has to be spent on writing scripts to deal with formatting issues. Additionally, we would welcome clearer guidelines from the Text Encoding Initiative on how structural and linguistic mark-up should be merged in practice.

### 6. Conclusion

With no single processing platform suitable for our needs and no clear set of guidelines to follow, identifying an adequate annotation format for our corpus has turned up a range of problems which had not been anticipated. Given the amount of investigation and pre- and post-processing work necessary to create a standardised annotation, it comes as no surprise that projects on a limited budget still prefer to use their own specialised formats. In the interest of interoperability and comparative studies between corpora we would welcome the development of clearer procedures whereby structural and linguistic annotations might be merged, and would wish to contribute actively to this process by sharing our experiences.

---


9<http://www.lemnitzer.de/lothar/KoLi/>
7. References


