A Gold Standard Corpus of Early Modern German

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

Abstract

This paper describes an annotated gold standard sample corpus of Early Modern German containing over 50,000 tokens of text manually annotated with POS tags, lemmas, and normalised spelling variants. The corpus is the first resource of its kind for this variant of German, and represents an ideal test bed for evaluating and adapting existing NLP tools on historical data. We describe the corpus format, annotation levels, and challenges, providing an example of the requirements and needs of smaller humanities-based corpus projects.

1 Introduction

This paper describes work which is part of a larger project whose goal is to develop a representative corpus of Early Modern German from 1650-1800. 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. One major goal is to annotate GerManC with linguistic information in terms of POS tags, lemmas, and normalised spelling variants. However, due to the lexical, morphological, syntactic, and graphemic peculiarities characteristic of Early Modern German, automatic annotation of the texts poses a major challenge. Most existing NLP tools are tuned to perform well on modern language data, but perform considerably worse on historical, non-standardised data (Rayson et al., 2007). This paper describes a gold standard subcorpus of GerManC which has been manually annotated by two human annotators for POS tags, lemmas, and normalised spelling variants. The corpus will be used to test and adapt modern NLP tools on historical data, and will be of interest to other current corpus-based projects in historical linguistics (Jurish, 2010; Fasshauer, 2011; Dipper, 2010).

2 Corpus design

2.1 GerManC

In order to enable corpus-linguistic investigations, the GerManC corpus aims to be representative on three different levels. First of all, the corpus includes a range of text types: four orally-oriented genres (dramas, newspapers, letters, and sermons), and four print-oriented ones (narrative prose, and humanities, scientific, and legal texts). Secondly, in order to enable historical developments to be traced, the period has been divided into three fifty year sections (1650-1700, 1700-1750, and 1750-1800). The combination of historical and text-type coverage should enable research on the evolution of style in different genres (cf. Biber and Finegan, 1989). Finally, the corpus also aims to be representative with respect to region, including five broad dialect areas: North German, West Central, East Central, West Upper (including Switzerland), and East Upper German (including Austria). Per genre, period, and region, three extracts of around 2000 words are selected, yielding a corpus size of nearly a million words. The structure of the GerManC corpus is summarised in Table 1.

2.2 GerManC-GS

In order to facilitate a thorough linguistic investigation of the data, the final version of the Ger-

Proceedings of the Fifth Law Workshop (LAW V), pages 124–128,
Portland, Oregon, 23-24 June 2011. ©2011 Association for Computational Linguistics

Periods	Regions	Genres
1650-1700	North	Drama
1700-1750	West Central	Newspaper
1750-1800	East Central	Letter
	West Upper	Sermon
	East Upper	Narrative
		Humanities
		Scientific
		Legal

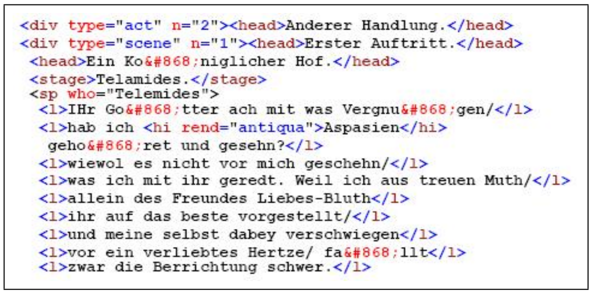
Table 1: Structure of the GerManC corpus

ManC corpus aims to provide the following linguistic annotations: 1.) Normalised spelling variants; 2.) Lemmas; 3.) POS tags. However, due to the non-standard nature of written Early Modern German, and the additional variation introduced by the three variables of 'genre', 'region', and 'time', automatic annotation of the corpus poses a major challenge. In order to assess the suitability of existing NLP tools on historical data, and with a view to adapting them to improve their performance, a manually annotated gold standard subcorpus has been developed, which aims to be as representative of the main corpus as possible (GerManC-GS). To remain manageable in terms of annotation times and cost, the subcorpus considers only two of the three corpus variables, 'genre' and 'time', as they alone were found to display as much if not more variation than 'region'. GerManC-GS thus only includes texts from the North German dialect region, with one sample file per genre and time period. Table 2 provides an overview of GerManC-GS, showing publication year, file name, and number of tokens for each genre/period combination. It contains 57,845 tokens in total, which have been manually annotated as described in the following sections.

2.3 Corpus format

As transcription 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, the raw input texts have been annotated according to the guidelines of the Text Encoding Initiative (TEI)[1] during manual transcription. The TEI have published a set of XML-based encoding conventions recommended for meta-textual markup to minimise inconsistencies across projects and to maximise mutual usability and data interchange.

The GerManC corpus has been marked up using the TEI P5 Lite tagset, which serves as standard for many humanities-based projects. Only the most relevant tags have been selected to keep the document structure as straightforward as possible. Figure 1 shows structural annotation of a drama excerpt, including headers, stage directions, speakers, as well as lines.



```
<div type="act" n="2"><head>Anderer Handlung.</head>
<div type="scene" n="1"><head>Erster Auftritt.</head>
 <head>Ein Ko&#868;niglicher Hof.</head>
 <stage>Telamides.</stage>
 <sp who="Telemides">
  <l>IHr Go&#868;tter ach mit was Vergnu&#868;gen/</l>
  <l>hab ich <hi rend="antiqua">Aspasien</hi>
   geho&#868;ret und gesehn?</l>
  <l>wiewol es nicht vor mich geschehn/</l>
  <l>was ich mit ihr geredt. Weil ich aus treuen Muth/</l>
  <l>allein des Freundes Liebes-Bluth</l>
  <l>ihr auf das beste vorgestellt/</l>
  <l>und meine selbst dabey verschwiegen</l>
  <l>vor ein verliebtes Hertze/ fa&#868;llt</l>
  <l>zwar die Berrichtung schwer.</l>
```

Figure 1: TEI annotation of raw corpus

3 Linguistic annotation

GerManC-GS has been annotated with linguistic information in terms of normalised word forms, lemmas, and POS tags. To reduce manual labour, a semi-automatic approach was chosen whose output was manually corrected by two trained annotators. The following paragraphs provide an overview of the annotation types and the main challenges encountered during annotation.

3.1 Tokenisation and sentence boundaries

As German orthography was not yet codified in the Early Modern period, word boundaries were difficult to determine at times. 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*). Our tokenisation scheme takes clitics into account, but does not yet deal with multi-word tokens. This means that whitespace characters usually act as token boundaries.

[1] http://www.tei-c.org

Genre	P	Year	File name	Tokens	Genre	P	Year	File name	Tokens
DRAM	**1**	1673	Leonilda	2933	**NARR**	**1**	1659	Herkules	2345
	2	1749	AlteJungfer	2835		**2**	1706	SatyrischerRoman	2379
	3	1767	Minna	3037		**3**	1790	AntonReiser	2551
HUMA	**1**	1667	Ratseburg	2563	**NEWS**	**1**	1666	Berlin1	1132
	2	1737	Königstein	2308		**2**	1735	Berlin	2273
	3	1772	Ursprung	2760		**3**	1786	Wolfenbuettel1	1506
LEGA	**1**	1673	BergOrdnung	2534	**SCIE**	**1**	1672	Prognosticis	2323
	2	1707	Reglement	2467		**2**	1734	Barometer	2438
	3	1757	Rostock	2414		**3**	1775	Chemie	2303
LETT	**1**	1672	Guericke	2473	**SERM**	**1**	1677	LeichSermon	2585
	2	1748	Borchward	2557		**2**	1730	JubelFeste	2523
	3	1798	Arndt	2314		**3**	1770	Gottesdienst	2292
								Total number of tokens	**57,845**

Table 2: GerManC-GS design

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 considerably across the corpus. For example, the virgule symbol “/” was often used in place of both comma and full-stop, which proves problematic for sentence boundary detection.

3.2 Normalising spelling variants and lemmatisation

One of the key challenges in working with historical texts is the large amount of spelling variation they contain. As most existing NLP tools (such as POS-taggers or parsers) are tuned to perform well on modern language data, they are not usually able to account for variable spelling, resulting in lower overall performance (Rayson et al., 2007). Likewise, modern search engines do not take spelling variation into account and are thus often unable to retrieve all occurrences of a given historical search word. Both issues have been addressed in previous work through the task of spelling normalisation. Ernst-Gerlach and Fuhr (2006) and Pilz and Luther (2009) have created a tool that can generate variant spellings for historical German to retrieve relevant instances of a given modern lemma, while Baron and Rayson (2008) and Jurish (2010) have implemented tools which normalise spelling variants in order to achieve better performance of NLP tools such as POS taggers (by running the tools on the normalised input). Our annotation of spelling variants aims to compromise between these two approaches by allowing for historically accurate linguistic searches, while also aiming to maximise the performance of automatic annotation tools. We treat the task of normalising spelling variation as a type of pre-lemmatisation, where each word token occurring in a text is labelled with a normalised head variant. As linguistic search requires a historically accurate treatment of spelling variation, our scheme has a preference for treating two seemingly similar tokens as separate items on historical grounds (e.g. *etwan* vs. *etwa*). However, the scheme normalises variants to a modernised form even where the given lexical item has since died out (e.g. obsolete verbs ending in *-iren* are normalised to *-ieren*), in order to support automatic tools using morphological strategies such as suffix probabilities (Schmid, 1994).

Lemmatisation resolves the normalised variant to a base lexeme in modern form, using Duden[2] pre-reform spelling. With obsolete words, the leading form in Grimm's Deutsches Wörterbuch[3] is taken.

3.3 POS-Tagging

We introduce a modified version of the STTS tagset (Schiller et al., 1999), the STTS-EMG tagset, to account for important differences between modern and Early Modern German (EMG), and to facilitate more accurate searches. The tagset merges two categories, as the criteria for distinguishing them are not applicable in EMG (1.), and provides a number of additional ones to account for special EMG constructions (2. to 6.):

[2] http://www.duden.de/

[3] http://www.dwb.uni-trier.de/

1. **PIAT** (merged with **PIDAT**): Indefinite determiner (occurring on its own, or in conjunction with another determiner), as in '***viele solche*** *Bemerkungen*'
2. **NA**: Adjectives used as nouns, as in '*der* ***Gesandte***'
3. **PAVREL**: Pronominal adverb used as relative, as in '*die Puppe,* ***damit*** *sie spielt*'
4. **PTKREL**: Indeclinable relative particle, as in '*die Fälle,* ***so*** *aus Schwachheit entstehen*'
5. **PWAVREL**: Interrogative adverb used as relative, as in '*der Zaun,* ***worüber*** *sie springt*'
6. **PWREL**: Interrogative pronoun used as relative, as in '*etwas,* ***was*** *er sieht*'

Around 2.0% (1132) of all tokens in the corpus have been tagged with one of the above POS categories, of which the merged PIAT class contains the majority (657 tokens). The remaining 475 cases occur as NA (291), or as one of the new relative markers PWAVREL (69), PWREL (57), PTKREL (38), and PAVREL (20).

4 Annotation procedure and agreement

In order to produce the gold standard annotations in GerManC-GS we used the GATE platform, which facilitates automatic as well as manual annotation (Cunningham et al, 2002). Initially, GATE's German Language plugin[4] was used to obtain word tokens and sentence boundaries. The output was manually inspected and corrected by one annotator, who manually added a layer of normalised spelling variants (NORM). This annotation layer was then used as input for the TreeTagger (Schmid, 1994), obtaining annotations in terms of lemmas (LEMMA) and POS tags (POS). All annotations (NORM, LEMMA, and POS) were subsequently corrected by two annotators, and all disagreements were reconciled to produce the gold standard. Table 3 shows the overall agreement for the three annotation types across GerManC-GS (measured in accuracy).

The agreement values demonstrate that normalised word forms and lemmas are relatively easy to determine for the annotators, with 96.9% and 95.5% agreement, respectively. POS tags, on the other, represent more of a challenge with only 91.6% agreement between two annotators, which is considerably lower than the agreement level reported for annotating a corpus of modern German using STTS, at 98.6% (Brants, 2000a). While a more detailed analysis of the results remains to be carried out, an initial study shows that POS agreement is lower in earlier texts (89.3% in Period P1) compared to later ones (93.1% in P3). It is likely that a substantial amount of disagreements in the earlier texts are due to the larger number of unfamiliar word forms and variants on the one hand, and foreign word tokens on the other. These represent a problem as from a modern view point it is not always easy to decide which words were 'foreign' to a language and which ones 'native'.

	NORM	**LEMMA**	**POS**
Agreed tokens (out of 57,845)	56,052	55,217	52,959
Accuracy (%)	**96.9%**	**95.5%**	**91.6%**

Table 3: Inter-annotator agreement

5 Future work

The gold standard corpus described in this paper will be used to test and adapt modern NLP tools on Early Modern German data. Initial experiments focus on utilising the layer of normalised spelling variants to improve tagger performance, and investigating to what extent normalisation can be reliably automated (Jurish, 2010). We further plan to retrain state-of-the-art POS taggers such as the TreeTagger and TnT Tagger (Brants, 2000b) on our data.

Finally, we plan to investigate how linguistic annotations can be automatically integrated in the TEI-annotated version of the corpus to produce TEI-conformant output. Currently, both structural and linguistic annotations are merged in GATE stand-off XML format, which, as a consequence, is no longer TEI-conformant. In the interest of interoperability and comparative studies between corpora we aim to contribute towards the development of clearer procedures whereby structural and linguistic annotations might be merged (Scheible et al., 2010).

[4] http://gate.ac.uk/sale/tao/splitch15.html

References

Alistair Baron and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics, Birmingham, UK*.

Douglas Biber and Edward Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language 65. 487-517*.

Torsten Brants. 2000a. Inter-annotator agreement for a German newspaper corpus. *Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece*.

Torsten Brants. 2000b. TnT – a statistical part-of-speech tagger. *Proceedings of the 6th Applied NLP Conference, ANLP-2000, Seattle, WA*.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Stefanie Dipper. 2010. POS-Tagging of historical language data: First experiments in semantic approaches in Natural Language Processing. *Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10). Saarbrücken, Germany. 117-121*.

Andrea Ernst-Gerlach and Norbert Fuhr. 2006. Generating search term variants for text collections with historic spellings. *Proceedings of the 28th European Conference on Information Retrieval Research (ECIR 2006), London, UK*.

Vera Fasshauer. 2011. http://www.indogermanistik.uni-jena.de/index.php?auswahl=184
Accessed 30/03/2011.

Bryan Jurish. 2010. Comparing canonicalizations of historical German text. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON), Uppsala, Sweden. 72-77*.

Thomas Pilz and Wolfram Luther. 2009. Automated support for evidence retrieval in documents with non-standard orthography. *The Fruits of Empirical Linguistics. Sam Featherston and Susanne Winkler (eds.). 211–228*.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. *Proceedings of the Corpus Linguistics Conference (CL2007), University of Birmingham, UK*.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2010. Annotating a Historical Corpus of German: A Case Study. *Proceedings of the LREC 2010 Workshop on Language Resources and Language Technology Standards, Valletta, Malta*.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. *Technical Report. Institut für maschinelle Sprachverarbeitung, Stuttgart*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing, Manchester, UK. 44–49*.