GATEtoGerManC: A GATE-based Annotation Pipeline for Historical German

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

University of Manchester
Oxford Road, Manchester, M13 9PL
scheible@ims.uni-stuttgart.de, richard.whitt@strath.ac.uk, {Martin.Durrell, Paul.Bennett}@manchester.ac.uk

Abstract

We describe a new GATE-based linguistic annotation pipeline for Early Modern German, which can be used to annotate historical texts with word tokens, sentence boundaries, lemmas, and POS tags. The pipeline is based on a customisation of the freely available ANNIE system for English (Cunningham et al., 2002), in combination with a version of the TreeTagger (Schmid, 1994) trained on gold standard Early Modern German data. The POS-tagging and lemmatisation components of the pipeline achieve an average accuracy of 89.44% and 83.16%, respectively, on unseen historical data from various genres and publication dates within the Early Modern period. We show that normalisation of spelling variation can further improve these results. With no specialised tools available for processing this particular stage of the language, this pipeline will be of particular interest to smaller, humanities-based projects wishing to add linguistic annotations to their historical data but which lack the means or resources to develop such tools themselves.

Keywords: historical; corpus; pipeline

1. Introduction

Research projects in the area of linguistics rely more and more on the use of digital text corpora for their investigations. In recent years, there has been a growing interest in studying historical language varieties using quantitative methods (Beal et al. (2007), Lindquist and Mair (2004)). In order to allow for more advanced corpus-linguistic investigations, it is desirable to annotate historical corpora with linguistic information such as lemmas and POS tags.

So far, few specialised tools are available for processing historical data, which is largely due to a lack of gold standard historical training data. Recently, a number of studies addressed the topic of POS-tagging historical data. For example, Dipper (2010) assessed the performance of a POS-tagger on Middle High German text, taking different levels of token normalisation and dialect subcorpora into account. For this purpose she trained and tested the tagger on historical word forms, with results ranging from 86% to 92% accuracy. Sánchez-Marco et al. (2011) report on an approach for Old Spanish where a modern POS-tagger's dictionary was expanded with historical variants and the tagger retrained on a small historical training corpus. They quote results of 94.5% accuracy for the main part of speech and 92.6% for lemmas on a gold standard dataset of 30,000 tokens (using 5-fold cross-validation). Further experiments were carried out for Old Norse where a tagger was trained on texts from the 13th/14th century, yielding 92.7% accuracy (Rögnvaldsson and Helgadóttir, 2008).

Recently, a number of projects have attempted to incorporate linguistic annotation tools auch as POS-taggers into a pipeline for processing historical data. Hinrichs and Zastrow (2012) report on an automated pipeline for a diachronic corpus of German (consisting of selected materials from the German Gutenberg Project), which adds different linguistic annotation layers to the corpus, including part-of-speech, lemmas, and constituent structure. The tools included in the pipeline were, however, trained on modern data (Tübinger Baumbank des Deutschen). Sánchez-Marco et al. (2011) incorporated their POS-tagger for Old Spanish into FreeLing, a library of language analysis services (Padró et al., 2010), along with other processing modules, such as a tokeniser. Their goal was to offer a complete set of tools for handling Old Spanish, which should be easily portable and reusable for other corpora and languages.

The present paper picks up on these ideas and describes a new text processing pipeline for annotating German texts from the Early Modern period (1650-1800), which has been developed on the GerManC corpus[1] and incorporated in the GATE text processing platform[2]. The GerManC corpus is a new corpus of Early Modern German covering the years 1650-1800. It 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 (Biber et al., 1994). GerManC includes eight different genres and is subdivided into three 50-year periods and the five major regions of the German Empire between 1650-1800. Like ARCHER, it consists of sample texts of 2,000 words (yielding around one million words altogether). The annotation pipeline (referred to as `GATEtoGerManC`) was originally developed to add linguistic information in terms of tokens, sentence boundaries, POS tags, and lemmas to structurally annotated texts. Its components are largely based on existing processing resources in GATE, which were extended and optimised for dealing with Early Modern German. One special feature of the pipeline is that it is document structure-aware, which means that it can utilise structural annotations present in the input documents for selecting and fine-tuning the linguistic annotations added to the data.

As the GerManC corpus displays such a wealth of variation, it lends itself as an ideal test bed for evaluating annotation tools such as POS taggers on Early Modern German data. For this purpose we prepared a substantial gold standard

[1] http://tinyurl.com/germanc
[2] www.gate.ac.uk

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

data set (consisting of almost 60,000 tokens annotated with POS tags and lemmas) which allows a fair evaluation of tools by assessing their robustness across different genres and sub-periods. The experiments described in this paper show that the performance of the POS-tagging and lemmatisation components in GATEtoGerManC are considerably better than the results achieved by modern tools on our data. The pipeline therefore promises to be of particular interest to researchers wishing to add linguistic annotations to historical German data, but who lack the means or resources to develop such tools themselves.

This paper aims to provide an overview of the GATEtoGerManC pipeline. For this purpose, we first introduce the development corpus GerManC (Section 2). Then, we describe the individual components of the pipeline, the tokeniser (Section 3.2), sentence splitter (3.3), and lemmatiser and POS-tagger (3.4). The final part of the paper shows that GATEtoGerManC can utilise structural mark-up of the input corpus (e.g. structural TEI annotations) in the annotation process (Section 4).

2. The GerManC corpus

2.1. Design

The GerManC corpus was compiled to enable corpus-linguistic investigations of the development and standardisation of German in the Early Modern period (1650-1800), and aims to be representative on three different levels. First of all, it includes a large variety 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, the period was divided into three fifty year sections (1650-1700, 1700-1750, and 1750-1800), which allows a study of historical developments over time. Finally, the corpus also takes regional variation into account by 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 2,000 words were selected, yielding a corpus size of nearly one million words. Table 1 summarises the structure of the GerManC corpus.

2.2. Gold standard subcorpus

The lemmatisation and POS-tagging components of the GATEtoGerManC pipeline were developed on a manually annotated gold standard subcorpus of GerManC, GerManC-GS. The subcorpus was created to allow an assessment of the suitability of existing NLP tools, with the goal of adapting them to improve their performance on historical data. For this reason, GerManC-GS aims to be as representative of the main corpus as possible. To remain manageable in terms of annotation times and cost, the subcorpus only includes texts for two of the three corpus variables, 'genre' and 'time', as they were found to display more variation than 'region'. GerManC-GS therefore includes one sample file per genre and time period from the North German region, resulting in 57,845 tokens in total. The subcorpus was annotated with gold standard POS tags, lemmas, and normalised word forms using a semi-automatic approach (Scheible et al., 2011a).

2.3. Structural annotations

The transcribed GerManC texts were annotated with structural information according to the guidelines of the Text Encoding Initiative (TEI)[3]. As GerManC is a diachronic corpus which will primarily be used by historical linguists, its annotation needs differ significantly from the large-scale corpora required in computational linguistics. For example, 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. The structural annotations in GerManC conform to the TEI P5 Lite tagset, which offers suitable strategies for encoding structural details such as found in our corpus. These vary considerably across the different genres, as illustrated in Figures 1, 2, and 3, which show three TEI-annotated excerpts from our corpus.

```
<head>
Schuel-Ordnung/
Fu℮r das Chur-Fu℮rstenthumb Bayern.</head>
<div type="section" n="1"><head>I.</head>
<head>Von Ambt/ vnd Belohnung der Schuelmeister.</head>
<p>DIe Schuelmeister sollen jhre vnderhabende Kinder
vorderist zu aller Christlichen Zucht/ Erbarkeit vnnd
Gottesforcht/ mit ho℮chstem Fleiß anhalten/ vnnd keines
 weegs zweifflen/ sie laisten hieran der Go℮ttlichen
Mayesta℮t einen sehr angenemmen Dienst/ so jhnen neben
jhrem gebu℮hrenden Schuel- oder QuatemberGelt/ welches
jedes OrthsObrigkeit zubestim¯en hat/ vbersichlich zu
seiner Zeit wird vergolten werden.</p></div>
```

Figure 1: TEI annotation of GerManC legal text

```
<div type="act" n="4">
<div type="scene" n="2"><head>Zweyte Scene.</head>
<head>Galerie im Schloß.</head>
<stage>Ra℮uber Moor. Amalia treten auf.</stage>
<sp who="Amalia"><speaker>Amalia.</speaker>
<p>Und getrauten Sie sich wol, sein Bildnis unter
diesen Gema℮lden zu erkennen?</p></sp>
<sp who="Moor"><speaker>Moor.</speaker>
<p>O ganz gewis. Sein Bild war immer lebendig in mir.
 <stage>(An den Gema℮lden herumgehend.)</stage><emph
rend="Sperrdruck">Dieser</emph> ists nicht.</p></sp>
```

Figure 2: TEI annotation of GerManC drama text

While the structure of the legal text in Figure 1 comprises common document structure mark-up such as headers ("head") and paragraph tags ("p"), the drama excerpt shown in Figure 2 includes more complex structural

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

```
<lg>
    <l>Gedenk' an deins Sohns bittern Tod/</l>
    <l>Sieh' an sein heil'ge Wunden roth/</l>
    <l>Die sind ja fu͛r die ganze Welt/</l>
    <l>Die Zahlung und das Lo͛segeld/</l>
    <l>Deß tro͛sten wir uns allezeit/</l>
    <l>Und hoffen auf Barmherzigkeit.</l>
</lg>
<p>Zuvo͛rderst erinnert sich ein wahrer Christ o͛ffters
 der unendlichen Liebe GOttes allen Menschen zu ihrem
Heil in Christo JEsa erwiesen mit Paulo/ der da sagt
<note place="margin-left"><bibl>1. Tim. 1, 13</bibl>
</note>: <quote>Das ist je gewißlich war/ und ein
theures werthes Wort/ daß JEsus Christus kommen ist in
die Welt/ die Su͛nder seelig zu machen.</quote></p>
```

Figure 3: TEI annotation of GerManC sermon text

mark-up in terms of headers (tag "head"), stage directions ("stage"), speakers (including co-reference, "sp" and "speaker"), and typeface changes (in this case, to indicate emphasis "emph"). The sermon excerpt (Figure 3) illustrates the use of line groups ("lg" and "l"), notes placed at the margin ("note"), bibliographical material ("bibl"), and quotes ("quote").

3. The `GATEtoGerManC` pipeline for Early Modern German

This section describes the linguistic annotation pipeline we developed in GATE (Cunningham et al., 2011) for annotating the data in our corpus. GATE ("General Architecture for Text Engineering") is an open source text engineering platform which is highly customisable and supports the following steps:

1. Loading XML-annotated corpora.
2. Implementing and running a linguistic annotation pipeline on the corpus documents.
3. Saving original TEI markup and new linguistic markup in GATE XML-standoff format.
4. Manual correction of the results.
5. Querying the resulting annotated corpora via the ANNIC Search GUI.

In this paper we focus on Step 2., which we addressed by implementing a linguistic annotation pipeline for Early Modern German (`GATEtoGerManC`). This pipeline is based on a customisation of the freely available ANNIE system for English ("A Nearly-New Information Extraction system") in GATE, which relies on finite state algorithms and the JAPE language (Cunningham et al., 2002). Using a platform such as GATE for implementing our pipeline is useful for a number of reasons. First of all, its stand-off annotation model is well-documented and accepted by the community, and allows original markup on the documents to be merged with newly created annotations. In addition, GATE offers methods for taking original markup into account when running processing tools, which means that the structural information provided by the TEI annotation layer can be utilised during linguistic annotation (cf. Section 4). GATE also incorporates facilities for manually correcting the output of the annotation pipeline, and for querying the results via the ANNIC search GUI.

The goal of the `GATEtoGerManC` pipeline is to add linguistic annotations in terms of word tokens, sentence boundaries, lemmas, and POS tags to the input documents. As discussed in Scheible et al. (2011b), each annotation component to be included in the pipeline requires careful consideration and adaptation as German orthography was not yet codified in the Early Modern period. Figure 4 provides an overview of the processing components of `GATEtoGerManC`, which we describe in the following subsections.

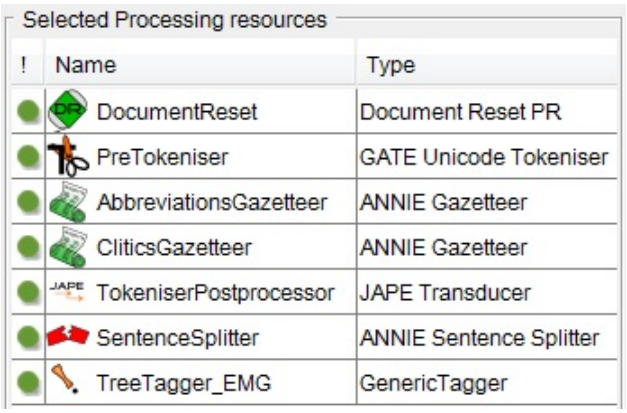

Figure 4: `GATEtoGerManC` components

3.1. Document reset

The document reset resource stems from the original ANNIE implementation and enables documents in the corpus to be reset to their original state. This resource is added to the beginning of the pipeline for cases where the pipeline has to be run several times. `DocumentReset` is set to remove all new annotation sets and their contents, but retaining the original TEI markup of the documents (referred to in GATE as "Original markups").

3.2. Tokenisation

`GATEtoGerManC`'s Tokenisation module consists of three main parts: a pre-tokeniser, two gazetteers, and a JAPE transducer, which adjusts the pre-tokeniser's output based on the information gathered through the gazetteers. The pre-tokeniser uses GATE's Unicode tokeniser resource to break up the input strings into simple initial tokens (`PreTokeniser` in Fig. 4). It is set to distinguish between numbers, punctuation, and words. A number of adaptations were made to the source code of the unicode tokeniser to handle Early Modern German (EMG) input more accurately. For example, the original tokeniser rules file does not account for typographic variants typically found in EMG, such as certain ligatures (where two or more graphemes are joined as a single glyph, as in Æ) or combining letters such as a superscripted `e` in place of an umlaut (as in o^{e}). To achieve accurate tokenisation, the appropriate unicode character classes were added to the tokenisation rules. Further rules were added to treat hyphenated compound nouns as one token rather than three, such as *Stadt-Kirche* ('town church'), *Slar-Affen* ('Cockaigne'), or *Aus-sicht* ('view'). These would also be treated

as single tokens according to modern German orthography (*Stadtkirche*, *Schlaraffen*, and *Aussicht*, respectively).

The second part of the tokenisation stage is the Gazetteer phase, where two gazetteers are consulted to mark 1.) potential abbreviations; and 2.) clitics. The first gazetteer, `AbbreviationsGazetteer`, contains a list of more than 800 potential EMG abbreviations. These were collected from GerManC, and had been either marked as abbreviations in the TEI markup, or discovered subsequently in other EMG data. Table 2 shows a number of examples from this list. The entries are case-sensitive, and the list may include several potential abbreviations of the same token. For example, *Holl*, *Holla^{e}nd*, or *Holla^{e}ndis* in Table 2 are all potential abbreviations of (inflected forms of) the adjective *Holländisch* ('Dutch').

Hanoveris	Hn	Hochsel	INTR
Hauptm	Hoch-Fu^{e}rstl	Holl	Ih
Heiligk	HochEdl	Holla^{e}nd	Ihr
Herrl	HochEhrw	Holla^{e}ndis	J
Herzogl	Hochfl	Hollsteinis	JHM
Heyl	Hochfu^{e}rstl	Hr	Ja^{e}n

Table 2: Examples of potential abbreviations

The second gazetteer aims to address the problem of tokenising clitics discussed in Scheible et al. (2011b). In EMG, clitics often occur in non-standard forms, such as *hastu*, a clitic version of *hast du* ('have you'), which should be tokenised as *has|tu*. The `CliticsGazetteer` aims to identify such cases in the input texts, and includes a "breakIndex" feature, which indicates the index of the token boundary, counting backwards starting at the end of the input token. For example, for *hastu*, the breakIndex feature is set to -2. This information is then utilised by the `TokeniserPostprocessor` during the next stage of the pipeline. There is also an alternative version of the `CliticsGazetteer`, which can be used to tokenise texts in which German '*to*-infinitive' verb forms are directly appended to the infinitival marker *zu* without intervening whitespace (e.g. *zubestellen* instead of *zu bestellen*, 'to order'; cf. Scheible et al. (2011b)). To allow for correct tokenisation as separate forms (*zu|bestellen*), `CliticsGazetteer2` includes an extensive list of such cases along with their appropriate "breakIndex" feature. The reason why '*to*-infinitive' clitics are not included in the main `CliticsGazetteer` is that they are often ambiguous with infinitive forms of particle verbs, which indeed represent one token (e.g particle verb *zumachen* 'to close' vs. '*to*-infinitive' clitic *zumachen* 'to do'). Table 3 shows some examples of clitics included in `CliticsGazetteer2`.

Both gazetteers store their findings as "Lookup" annotations in the set of new annotations in GATE, and can be easily adapted and extended by adding new potential abbreviations and clitics to the respective lists.

The final stage of tokenisation is carried out by the JAPE transducer `TokeniserPostprocessor`. JAPE transducers are processing resources in GATE which consist of a grammar (stored in a .jape file) which defines changes to annotations in the document. The `TokeniserPostprocessor` resource contains a variety of rules to adjust the tokeniser output. For example, if a potential abbreviation identified by `AbbreviationsGazetteer` is followed by a full stop or colon token, the tokens are combined and marked as single token of the kind "abbreviation". The output of the `CliticsGazetteer` is dealt with by separating tokens marked as clitics according to their respective "breakIndex" attribute. For instance, the token *hastu* ('have you'), whose breakIndex feature is set to -2, is split into two separate tokens *has|tu*. Further rules were devised to join tokens such as *steh'n* ('stand') or *g'storben* ('died'), where the apostrophe usually indicates an omitted vowel, a common feature of colloquial speech (standard German: *stehen* and *gestorben*, respectively). Finally, the `TokeniserPostprocessor` also contains a number of rules adapted from the original ANNIE system in GATE, such as various rules for joining numbers.

zulassen	zuerfordern	zuschu^{e}tten
zuhalten	zuschlagen	zuwerffen
zumachen	zulegen	zuschmeltzen
zuhaben	zuwaschen	zuziehen
zuermahnen	zuverschlemmen	zugewarten

Table 3: Examples of potential '*to*-infinitive' clitics

3.3. Sentence Splitting

The next stage of the pipeline aims to mark up sentence boundaries. The GerManC sentence splitter is based on an adaptation of the ANNIE sentence splitter for English (Cunningham et al., 2002), which defines sentence splitting rules based on puctuation symbols such as full stops ".", question marks "?", and exclamation marks "!". In contrast to ANNIE, full stops indicating abbreviations rather than sentence splits are already taken care of in the tokenisation module of `GATEtoGerManC`. In ANNIE, abbreviations are only determined during sentence boundary detection.

One of the most problematic issues concerning sentence boundary detection in Early Modern German is that punctuation is not standardised and varies considerably across texts. Conventional modern markers of sentence boundaries which are included in the ANNIE sentence splitting rules (such as full stops, exclamation marks, and question marks) sometimes do not occur at all in historical texts. Instead, semi-colons, colons, and the virgule symbol "/" may have the function of marking both clause and sentence boundaries, and it is often difficult to decide which function was intended by the author. As this kind of variation is difficult to handle automatically, the `SentenceSplitter` takes the following general approach:

1. Semi-colons and colons are added to the punctuation list indicating sentence splits.
2. Virgule symbols are excluded.

This procedure aims to ensure that texts are divided up into

useful chunks. However, it is important to be aware of the fact that semi-colons and colons may sometimes indicate clause boundaries, or even lists, rather than full sentences. The virgule symbol, on the other hand, is not included in the list of sentence splitters. Even though it is sometimes used to mark sentence boundaries, it is used in place of a modern comma in most cases, and thus not marking a sentence split. Finally, while the ANNIE sentence splitter generally considers new lines as sentence boundaries, this is not accurate for data such as found in our corpus. Texts from the drama genre in particular often contain stanzas with sentences spanning several lines. The GATEtoGerManC sentence splitter contains rules which allow for sentences spanning several lines, as shown in Figure 5.

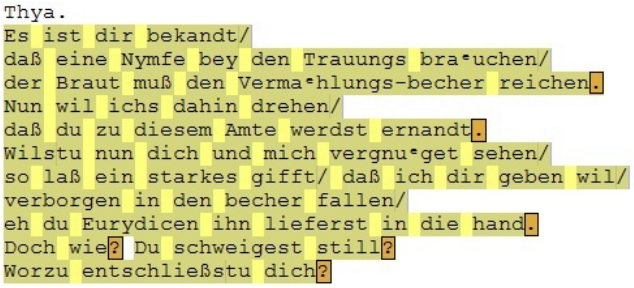

Figure 5: Annotation of sentence splits in a GerManC drama text (marked by black squares)

3.4. Lemmatisation and POS-Tagging

The final component of GATEtoGerManC adds token-based annotations in terms of lemmas and POS tags. Our lemmatisation scheme aims to resolve each token in the corpus to a base lexeme in modern form, using Duden[4] pre-reform spelling. With obsolete words, the leading form in Grimm's Deutsches Wörterbuch[5] is taken. The POS tagging scheme is based on the STTS tagset for German (Schiller et al., 1999), with a number of modifications to account for differences between modern and Early Modern German (EMG), and to facilitate more accurate searches. The STTS-EMG tagset thus contains a number of additional categories to account for special EMG constructions, such as various kinds of non-standard relative markers (Scheible et al., 2011b). The new POS categories account for around 2.0% of all tokens in the Gold Standard subcorpus of GerManC (cf. Section 2.2).

To add lemmatisation and POS-tagging to GATEtoGerManC, the GATE wrapper for the TreeTagger (Schmid, 1994) is added to the end of the pipeline (part of the TaggerFramework plugin)[6]. The TreeTagger is a probabilistic POS-tagger which uses decision trees to determine the appropriate size of context needed for estimating transition probabilities. It can be trained on any language, as long as a suitable lexicon and a manually tagged training corpus are available.

Previous experiments showed that using the original parameter files for modern German supplied with the tagger[7] only achieves moderate results for tagging EMG data, as reported in Scheible et al. (2011b), with an overall accuracy of only 69.6% on the gold standard test corpus described in Section 2.2. Our initial experiments further showed that normalisation of spelling variation could improve the results of the modern tagger by 10% (79.7% accuracy). Similar results were reported by Rayson et al. (2007) for English. Spelling variation is a well-known problem in the automatic processing of older language varieties. Non-standard spellings are particularly frequent in earlier texts in the GerManC corpus (35-40% of all tokens at the beginning of the early modern period, ca. 1650), while the proportion is lower in later texts (5-10% towards the end, ca. 1800).

To maximise the performance of the tagger for EMG data, we retrained it on our gold standard subcorpus GerManC-GS. Three models of the tagger are available: One trained on the original word forms in the corpus (EMG-ORIG), one trained on the gold standard normalised word forms (EMG-NORM$_{GS}$), and the third model was trained on normalised spelling variants produced by an automatic tool developed by Jurish (2010) (EMG-NORM$_A$). In all three cases, tagging results were further improved by merging the tagger's lexicon compiled from the training data with a modern lexicon derived from the TIGER corpus[8].

All tagger models were evaluated on the gold standard corpus using 'leave-one-out' cross-validation, where the 24 gold standard corpus files were used to carry out 24 train-and-test cycles, in which 23 files were used as training material, and the remaining one file for testing. Traditionally, POS-taggers are evaluated using *k*-fold cross validation, where the sentences in the corpus are randomly divided into k (usually 10) mutually exclusive partitions of approximately equal size, resulting in 10 train-and-test cycles. However, this method often provides an overoptimistic estimate of performance, as the training and test data are usually drawn from the same text type (or even from the same documents), and therefore tend to be very similar (Giesbrecht and Evert, 2009). The 'leave-one-out' error estimation avoids such over-fitting to the data, and offers a more accurate estimate of the performance of the tagger on unseen data. Our gold standard data is especially suitable for this evaluation technique due to its special structure, consisting of samples of equal size (ca. 2,000 words) drawn from eight different genres and three time periods.

Table 4 summarises the performance of the three models tested on 1.) the original word forms in the gold standard corpus (ORIG), and 2.) the output of the automatic normalisation tool (NORM$_A$). The results show that all retrained versions of the TreeTagger outperform the modern version (which only achieved 69.6% accuracy on the same data.)

The final version of the GATEtoGerManC pipeline includes the parameter files of all EMG tagging models. In addition, we provide a number of XSLT stylesheets which allow tokens to be exported for GATE-external processing,

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

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

[6] We would like to thank Mark Greenwood from the GATE team for his assistance in setting up the TreeTagger.

[7] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[8] Kindly provided by Stefanie Dipper, cf. Dipper (2011).

Model	Test data	Average accuracy	
		POS	LEMMA
EMG-ORIG	ORIG	89.44	83.16
EMG-ORIG	$NORM_A$	89.58	86.3
EMG-$NORM_A$	ORIG	85.66	77.57
EMG-$NORM_A$	$NORM_A$	89.81	86.83
EMG-$NORM_{GS}$	ORIG	80.25	77.1
EMG-$NORM_{GS}$	$NORM_A$	84.33	85.96

Table 4: Performance of TreeTagger trained on original (EMG-ORIG), manually normalised (EMG-$NORM_{GS}$) and automatically normalised data (EMG-$NORM_A$) in GerManC-GS

and to be re-imported by adding any new annotations to the list of token features (using GATE stand-off format). This is useful for incorporating the output of tools for which no GATE wrapper yet exists, or which are subject to copyright restrictions. We implemented these stylesheets to be able to incorporate the normalised spelling variants produced by Jurish (2010).

Figure 6 shows a screenshot of a drama text processed with `GATEtoGerManC`, showing token-based markup in terms of POS-tags (feature "pos"), lemmas (feature "lemma"), and normalised spelling variants (feature "norm").

Figure 6: Token-based markup produced by `GATEtoGerManC`

4. Making `GATEtoGerManC` document-structure-aware

In a recent paper, Poesio et al. (2011) note that while freely available HLT pipelines such as LingPipe, OpenNLP, or GATE support a variety of document formats as input, actual processing rarely takes advantage of structural information. They suggest that making pipelines document structure-aware can improve the overall annotation process, for example by distinguishing between titles and paragraph text, where the syntactic conventions are known to differ greatly. Furthermore, linguistic processing may only be useful for specific parts of a document (e.g. excluding bibliography sections of articles). We pick up on this idea by getting `GATEtoGerManC` to utilise the structural TEI annotation of the input corpus (cf. Section 2.3). This is done by using GATE's "Segment Processing" resource: This resource allows documents to be processed step-by-step by specifying a controller (i.e. a pipeline or processing resource) and the annotation segments the controller should be applied to. `GATEtoGerManC` contains a number of segment processors which are tailored towards the various genres included in the GerManC corpus, as each of the genres contains different kinds of structural markup. For example, the segment processor of the drama corpus allows the main `GATEtoGerManC` pipeline to be applied to segments marked as "speech" in the TEI-annotated version of the corpus, while speaker names, headers, and stage directions can be excluded from consideration (marked as "speaker", "head" and "stage" respectively in TEI). This can, for example, be useful for corpus-linguistic investigations which tend to focus on drama texts as an orally-oriented genre, where statistics on the use of certain linguistic features may be skewed by the extensive use of stage directions and speaker turns. Figure 7 illustrates the result of using the drama segment processor on the TEI excerpt shown in Figure 2.

```
Zweyte Scene.
Galerie im Schloß.
Ra*uber Moor. Amalia treten auf.
Amalia.
Und getrauten Sie sich wol, sein Bildnis
unter diesen Gema*lden zu erkennen?
Moor.
O ganz gewis. Sein Bild war immer
lebendig in mir. (An den Gema*lden
herumgehend.) Dieser ists nicht.
```

Figure 7: Example of drama segment processing in `GATEtoGerManC` (yellow markup = sentence, blue markup = tokens)

Similary, bibliographical information (marked as "bibl" in TEI) can be excluded from linguistic processing. Bibliographical references often interrupt the text flow. For example, the sermon subcorpus of GerManC contains many bible citations of the form "book chapter, verse(s)", as shown in Figure 3. To address this problem, `GATEtoGerManC` contains a segment processor which ignores text marked up as "bibl", resulting in the annotated version shown in Figure 8.

```
Zuvo*rderst erinnert sich ein wahrer Christ
o*ffters der unendlichen Liebe
GOttes allen Menschen zu ihrem Heil in
Christo JEsa erwiesen mit Paulo/ der da sagt
1. Tim. 1, 13: Das ist je gewißlich war/ und
ein theures werthes Wort/ daß JEsus Christus
```

Figure 8: Example of sermon segment processing in `GATEtoGerManC`

5. Conclusion

This paper describes a new GATE-based linguistic annotation pipeline for Early Modern German, which can be used to annotate historical texts with word tokens, sentence boundaries, lemmas, and POS tags. Its output is stored in GATE stand-off format and combines both structural and linguistic markup, which can be queried simultaneously using the ANNIC Search GUI incorporated in GATE. `GATEtoGerManC` is straightforward to use and adapt, and thus promises to be of interest for other researchers working on corpus-projects in historical German linguistics.

6. References

Joan C. Beal, Karen P. Corrigan, and Herrmann L. Moisl. 2007. *Creating and digitizing language corpora: Diachronic databases*. Palgrave, Houndmills.

Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994. ARCHER and its challenges: compiling and exploring A Representative Corpus of Historical English Registers. In Udo Fries, Peter Schneider, and Gunnel Tottie, editors, *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora*, pages 1–13.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*.

Stefanie Dipper. 2010. POS-tagging of historical language data: First experiments. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10)*, pages 117–121, Saarbrücken, Germany.

Stefanie Dipper. 2011. Morphological and Part-of-Speech tagging of historical language data: A comparison. *Journal for Language Technology and Computational Linguistics*, 26(2):25–37.

Eugenie Giesbrecht and Stefan Evert. 2009. Is Part-of-Speech tagging a solved task? An evaluation of POS taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.

Erhard Hinrichs and Thomas Zastrow. 2012. Linguistic annotations for a diachronic corpus of German. *Linguistic Issues in Language Technology*, 7.

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

Hans Lindquist and Christian Mair. 2004. *Corpus approaches to grammaticalization in English*. Benjamins, Amsterdam.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. FreeLing 2.1: Five years of open-source language processing tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, Valletta, Malta.

Massimo Poesio, Eduard Barbu, Egon Stemle, and Christian Girardi. 2011. Structure-preserving pipelines for digital libraries. In *Proceedings of the ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, Portland, Oregon.

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. In Matthew Davies, Paul Rayson, Susan Hunston, and Pernilla Danielsson, editors, *Proceedings of the Corpus Linguistics Conference (CL2007), University of Birmingham, UK*.

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2008. Morphological tagging of Old Norse texts and its use in studying syntactic variation and change. In *Proceedings of the LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakech, Morocco.

Cristina Sánchez-Marco, Gemma Boleda, and Lluís Padró. 2011. Extending the tool, or how to annotate historical language varieties. In *Proceedings of the ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, Portland, Oregon.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011a. A Gold Standard Corpus of Early Modern German. In *Proceedings of the ACL-HLT 2011 Linguistic Annotation Workshop (LAW V)*, pages 124–128, Portland, Oregon.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011b. Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *Proceedings of the ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, Portland, Oregon.

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. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.