Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text

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

The goal of this study is to evaluate an 'off-the-shelf' POS-tagger for modern German on historical data from the Early Modern period (1650-1800). With no specialised tagger available for this particular stage of the language, our findings will be of particular interest to smaller, humanities-based projects wishing to add POS annotations to their historical data but which lack the means or resources to train a POS tagger themselves. Our study assesses the effects of spelling variation on the performance of the tagger, and investigates to what extent tagger performance can be improved by using 'normalised' input, where spelling variants in the corpus are standardised to a modern form. Our findings show that adding such a normalisation layer improves tagger performance considerably.

1 Introduction

The work described in this paper is part of a larger investigation whose goal is to create a representative corpus of Early Modern German from 1650-1800. The GerManC corpus, which is due to be completed this summer, was developed to allow for comparative studies of the development and standardisation of English and German in the 17th and 18th centuries. In order to facilitate corpus-linguistic investigations, one of the major goals of the project is to annotate the corpus with POS tags. However, no specialised tools are yet available for processing data from this period. The goal of this study is therefore to evaluate the performance of an 'off-the-shelf' POS-tagger for modern German on data from the Early Modern period, in order to assess if modern tools are suitable for a semi-automatic approach, and how much manual post-processing work would be necessary to obtain gold standard POS annotations.

We report on our results of running the TreeTagger (Schmid, 1994) on a subcorpus of GerManC containing over 50,000 tokens of text annotated with gold standard POS tags. This subcorpus is the first resource of its kind for this variant of German, and due to its complex structure it represents an ideal test bed for evaluating and adapting existing NLP tools on data from the Early Modern period. The study described in this paper represents a first step towards this goal. Furthermore, as spelling variants in our corpus have been manually normalised to a modern standard, this paper also aims to explore the extent to which tagger performance is affected by spelling variation, and to what degree performance can be improved by using 'normalised' input. Our findings promise to be of considerable interest to other current corpus-based projects of earlier periods of German (Jurish, 2010; Fasshauer, 2011; Dipper, 2010). Before presenting the results in Section 4, we describe the corpus design (Section 2), and the preprocessing steps necessary to create the gold standard annotations, including adaptations to the POS tagset (Section 3).

2 Corpus design

In order to be as representative of Early Modern German as possible, the GerManC corpus design considers three different levels. First, the corpus includes a range of text types: four orally-oriented

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 19–23,
Portland, OR, USA, 24 June 2011.

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 is divided into three fifty year sections (1650-1700, 1700-1750, and 1750-1800). Finally, the corpus also aims to be representative with respect to region, including five broad areas: North German, West Central, East Central, West Upper (including Switzerland), and East Upper German (including Austria). Three extracts of around 2000 words were selected per genre, period, and region, yielding a corpus size of nearly a million words.

The experiments described in this paper were carried out on a manually annotated gold standard subcorpus of GerManC, GerManC-GS. The subcorpus was developed to enable an assessment of the suitability of existing NLP tools on historical data, with a view to adapting them to improve their performance. For this reason, GerManC-GS aims to be as representative of the main corpus as possible. However, to remain manageable in terms of annotation times and cost, the subcorpus only considers 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 includes texts from the North German region, with one sample file per genre and time period. The corpus contains 57,845 tokens in total, and was annotated with gold standard POS tags, lemmas, and normalised word forms (Scheible et al., to appear).

3 Creating the gold standard annotations

This section provides an overview of the preprocessing work necessary to obtain the gold standard annotations in GerManC-GS. We used the GATE platform to produce the initial annotations, which facilitates automatic as well as manual annotation (Cunningham et al., 2002). First, GATE's German Language plugin[1] was used to obtain word tokens and sentence boundaries. The output was manually inspected and corrected by one annotator, who further added a layer of normalised spelling variants. This annotation layer was then used as input for the TreeTagger (Schmid, 1994), obtaining annotations in terms of POS tags and lemmas. All annotations were subsequently corrected by two annotators, and disagreements were reconciled to produce the gold standard.

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

3.1 Tokenisation

As German orthography was not yet codified in the Early Modern period, a number of specific decisions had to be made in respect of tokenisation. For example, clitics can occur in various non-standard forms. To allow for accurate POS tagging, clitics should be tokenised as separate items, similar to the negative particle *n't* in *can't* in English, which is conventionally tokenised as *ca|n't*. A case in point is *hastu*, a clitic version of *hast du* ('have you'), which we tokenise as *has|tu*. Furthermore, German '*to*-infinitive' verb forms are often directly appended to the infinitival marker *zu* without intervening whitespace (e.g. *zugehen* instead of *zu gehen*, 'to go'). Such cases are tokenised as separate forms (*zu|gehen*) to allow for their accurate tagging as *zu*/PTKZU *gehen*/VVINF.

A further problem can be found in multi-word tokens, where the same expression is sometimes treated as a compound (e.g. *obgleich*), but at other times written separately (*ob gleich*). Such cases represent a problem for POS-tagging as the variants have to be treated differently even though their function in the sentence is the same. Our tokenisation scheme deals with these in a similar way to normal conjunctions consisting of two words, where the most suitable tags are assigned to each token (e.g. *als*/KOKOM *wenn*/KOUS). Thus, the compound *obgleich* is tagged KOUS, while the multiword variant *ob gleich* is tagged as *ob*/KOUS *gleich*/ADV.

3.2 Normalising spelling variants

All spelling variants in GerManC-GS were normalised to a modern standard. We view the task of normalising spelling variation as a type of prelemmatisation, where each word token occurring in a text is labelled with a normalised head variant. As linguistic searches require 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*). On the other hand, 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). Inter-annotator agreement for annotating spelling variation was 96.9%, which indicates that normalisation is a relatively easy task.

Figure 1 shows the proportion of normalised word tokens in the individual corpus files plotted against time. The graph clearly shows a decline of spelling variants over time: while the earlier texts contain 35-40% of normalised tokens, the proportion is lower in later texts (11.3% in 1790, and 5.4% in 1798). This suggests that by the end of the period (1800) codification of the German language was already at an advanced stage.

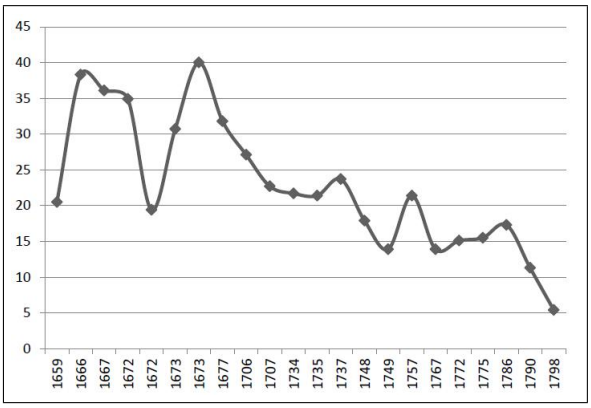

Figure 1: Proportion of normalised tokens (plotted against time)

3.3 Adapting the POS tagset (STTS)

To account for important differences between modern and Early Modern German (EMG), and to facilitate more accurate searches, we adapted the STTS tagset (Schiller et al., 1999). The STTS-EMG 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.):

1. **PIAT** (merged with **PIDAT**): Indefinite determiner, as in '***viele solche*** *Bemerkungen*' ('*many such remarks*')
2. **NA**: Adjectives used as nouns, as in '*der* ***Gesandte***' ('*the ambassador*')
3. **PAVREL**: Pronominal adverb used as relative, as in '*die Puppe,* ***damit*** *sie spielt*' ('*the doll with which she plays*')
4. **PTKREL**: Indeclinable relative particle, as in '*die Fälle,* ***so*** *aus Schwachheit entstehen*' ('*the cases which arise from weakness*')
5. **PWAVREL**: Interrogative adverb used as relative, as in '*der Zaun,* ***worüber*** *sie springt*' ('*the fence over which she jumps*')
6. **PWREL**: Interrogative pronoun used as relative, as in '*etwas,* ***was*** *er sieht*' ('*something which he sees*')

Around 2.0% (1132) of all tokens in the corpus were tagged with one of the above POS categories. Inter-annotator agreement for the POS tagging task was 91.6%.

4 'Off-the-shelf' tagger evaluation on Early Modern German data

The evaluation described in this section aims to complement the findings of Rayson et al. (2007) for Early Modern English, and a recent study by Dipper (2010), in which the TreeTagger is applied to a corpus of texts from Middle High German (MHG) - i.e. a period earlier than ours, from 1050-1350. Both studies report considerable improvement of POS-tagging accuracy on normalised data. However, unlike Dipper (2010), whose experiments involve retraining the TreeTagger on a modified version of STTS, our experiments assess the "off-the-shelf" performance of the modern tagger on historical data. We further explore the question of what effect spelling variation has on the performance of a tagger, and what improvement can be achieved when running the tool on normalised data.

Table 1 shows the results of running the TreeTagger on the original data vs. normalised data in our corpus using the parameter file for modern German supplied with the tagger[2]. The results show that while overall accuracy for running the tagger on the original input is relatively low at 69.6%, using the normalised tokens as input results in an overall improvement of 10% (79.7%).

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

	O	N
Accuracy	69.6%	79.7%

Table 1: TreeTagger accuracy on original (O) vs. normalised (N) input

However, improvement through normalisation is not distributed evenly across the corpus. Figure 2 shows the performance curves of using TreeTagger on original (O) and normalised (N) input plotted against publication date. While both curves gradually rise over time, the improvement curve (measured as difference in accuracy between N and O) diminishes, a direct result of spelling variation being more prominent in earlier texts (cf. Figure 1).

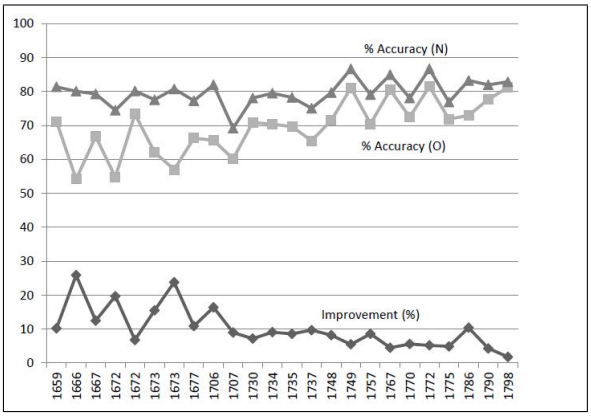

Figure 2: Tagger performance plotted against publication date

Compared with the performance of the TreeTagger on modern data (ca. 97%; Schmid, (1995)), the current results seem relatively low. However, two issues should be taken into account when interpreting these findings: First, the modern accuracy figures result from an evaluation of the tagger on the text type it was developed on (newspaper text), while GerManC-GS includes a variety of genres, which is bound to result in lower performance. Secondly, inter-annotator agreement was also found to be considerably lower in the present task (91.6%) than in one reported for modern German (98.6%; Brants, 2000a). This is likely to be due to the large number of unfamiliar word forms and variants in the corpus, which represent a problem for human annotators.

Finally, Figure 3 provides a more detailed overview of the effects of spelling variation on POS tagger performance. Of 12,744 normalised tokens in the corpus, almost half (5981; 47%) are only tagged correctly when using the normalised variants as input. Using the original word form as input results in a false POS tag in these cases. Overall, this accounts for an improvement of around 10.3% (5981 out of 57,845 tokens in the corpus). However, 32% (4119) of normalised tokens are tagged correctly using both N and O input, while 18% (2339) of tokens are tagged incorrectly using both types of input. This means that for 50% of all annotated spelling variants, normalisation has no effect on POS tagger performance. In a minority of cases (305; 3%) normalisation has a negative effect on tagger accuracy.

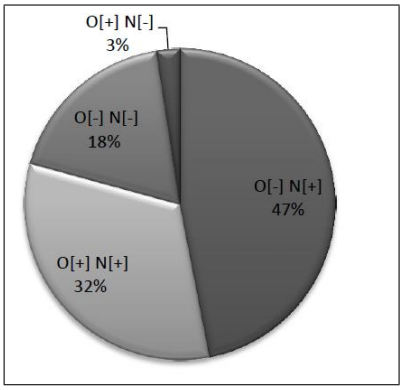

Figure 3: Effect of using original (O)/normalised (N) input on tagger accuracy for normalised tokens (+: correctly tagged; -: incorrectly tagged)

5 Conclusion and future work

The results of our study show that using an 'off-the shelf' German POS tagger on data from the Early Modern period achieves reasonable results (69.6% on average), but requires a substantial amount of manual post-editing. We further demonstrated that adding a normalisation layer can improve results by 10%. However, using the current manual normalisation scheme only half of all annotations carried out have a positive effect on tagger performance. In future work we plan to investigate if the scheme can be adapted to account for more cases, and to what extent normalisation can be reliably automated (Jurish, 2010). Finally, we plan to retrain state-of-the-art POS taggers such as the TreeTagger and TnT Tagger (Brants, 2000b) on our data and compare the results to the findings of this study.

References

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*.

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*.

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. To appear. A Gold Standard Corpus of Early Modern German. *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V), 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. *International Conference on New Methods in Language Processing, Manchester, UK. 44–49*.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop. 47–50*.