POS-Tagging of Historical Language Data: First Experiments

Stefanie Dipper
Institute of Linguistics
Ruhr-University Bochum
Germany
dipper@linguistics.rub.de

Abstract

This paper deals with part-of-speech tagging applied to manuscripts written in Middle High German. We present the results of a set of experiments that involve different levels of token normalization and dialect-specific subcorpora. As expected, tagging with "normalized", quasi-standardized tokens performs best (accuracy $> 91\%$). Training on slightly simplified word forms or on larger corpora of heterogeneous texts does not result in considerable improvement.

1 Introduction[1]

This paper deals with automatic analysis of historical language data, namely part-of-speech (POS) tagging of texts from Middle High German (1050–1350). Analysis of historical languages differs from that of modern languages in two important points.

First, there are no agreed-upon, standardized writing conventions. Instead, characters and symbols used by the writer of some manuscript in parts reflect impacts as different as spatial constraints (parchment is expensive, hence, use of abbreviations seems favorable) or dialect influences (the dialect spoken by the author of the text, or the writer's dialect, who writes up or copies the text, or even the dialect spoken by the expected readership). This often leads to inconsistent spellings, even within one text written up by one writer.

Second, resources of historical languages are scarce and often not very voluminous, and manuscripts are frequently incomplete or damaged.

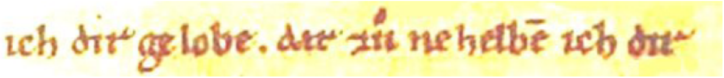
These features—data variance and lack of large resources—challenge many analysis tools, whose quality usually depend on the availability of large training samples. "Modern" POS taggers have been used mainly for the annotation of English historical corpora. The "Penn-Helsinki Parsed Corpora of Historical English" (Kroch and Taylor, 2000; Kroch et al., 2004) have been annotated in a bootstrapping approach, which involves successive cycles of manual annotation, training, automatic tagging, followed by manual corrections, etc. The project "GermanC"[2] uses a state-of-the-art tagger whose lexicon has been filled with historical form variants. In contrast, Rayson et al. (2007) and Pilz et al. (2006) automatically map historical word forms to the corresponding modern word forms, and analyze these by state-of-the-art taggers. The mappings make use of the Soundex algorithm, Edit Distance, or heuristic rules. Rayson et al. (2007) apply this technique for POS tagging, Pilz et al. (2006) for a search engine for texts without standardized spelling.

This paper reports on preliminary experiments in applying a state-of-the-art POS tagger (the TreeTagger, Schmid (1994)) to a corpus of texts from Middle High German (MHG). Our approach is similar to the one by Kroch et al. in that we train and apply the tagger to historical rather than modern word forms. Our tagging experiments make use of a balanced MHG corpus that is created and annotated in the context of the projects "Mittelhochdeutsche Grammatik" and "Referenzkorpus Mittelhochdeutsch".[3] The corpus has been semi-automatically annotated with POS tags, morphol-

[1] We would like to thank the anonymous reviewers for helpful comments. The research reported here was supported by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/1-1.

[2] http://www.llc.manchester.ac.uk/research/projects/germanc/

[3] http://www.mittelhochdeutsche-grammatik.de, http://www.linguistics.rub.de/mhd/



UNICODE	ich	dir	gelobe	.	dar	zuͦ	ne	helbē	ich	dir
STRICT	ich	dir	gelobe	.	dar	zu\o	ne	helbe\-	ich	dir
SIMPLE	ich	dir	gelobe	.	dar	zuo	ne	helben	ich	dir
NORM	ich	dir	gelobe	.	dar	zuo	ne	hilfen	ich	dir
LEMMA	ich	dû	ge-loben		dâr	zuo	ne	hëlfen	ich	dû
STTS	PPER	PPER	VVFIN	$.	ADV	ADV	PTK-NEG	VVFIN	PPER	PPER

Figure 1: A small excerpt from *Tristrant* (Magdeburg fragment), along with different types of transcriptions and annotations (screenshot from `http://www.hs-augsburg.de/~harsch/germanica/Chronologie/12Jh/Eilhart/eil_tmma.html`)

ogy, lemma, and a normalized word form, which represents a virtual historical standardized form. The corpus is not annotated with modern word forms.

In this paper, we present the results of a set of experiments that involve different types of tokens (simplified and normalized versions) and dialect-specific subcorpora. Sec. 2 presents detailed information about the corpus and its annotations, Sec. 3 addresses the tagging experiments and results.

2 The Corpus

The corpus is a collection of texts from the 12th–14th centuries, including religious as well as profane texts, prose and verse. The texts have been selected in a way as to cover the period of MHG as optimal as possible. The texts distribute in time, i.e. over the relevant centuries, and in space, coming from a variety of Upper German (UG) and Middle German (MG) dialects. UG dialects were (and are still) spoken in the Southern part of Germany, Switzerland, and Austria. Examples are Bavarian, Swabian, or Alemannic. MG dialects were spoken in the middle part of Germany, e.g. Franconian or Thuringian.

The texts are *diplomatic transcriptions*, i.e., they aim at reproducing a large range of features of the original manuscript or print, such as large initials, or variant letter forms (e.g. short vs. long s: <s> vs. <ſ>), or abbreviations (e.g. the superscribed hook <ʾ> can abbreviate *-er*: *cleidʾ* stands for *cleider* 'clothes'[4]).

The original corpus provides two different versions of "word forms": the diplomatic transcription and a normalized form. For the tagging experiments, we created simplified versions of these forms: "strict" and "simple" transcriptions, and a normalized ASCII-form. In the following, we describe both the original and simplified versions of the word forms. Figure 1 presents an example fragment encoded in the different versions.

- The *Unicode* version is the diplomatic transcription, as produced by human transcribers. It renders the manuscript very closely and, e.g., distinguishes short vs. long s, abbreviations, etc.
- The *strict* version is a slightly modified version of the original transcription which uses ASCII characters only. Instead of letters with diacritics or superposed characters (*ö, uͦ*), it uses ASCII characters combined with the backslash as an escape character (*o\", u\o*). Ligatures (*æ*) are marked by an underscore (*a_e*), *&* is mapped to *e_t*, *þ* to *t_h*.
- The *simple* version abstracts away from many of the diplomatic and/or dialectal idiosyncrasies. Characters are mapped to lower case, all kinds of accents or other diacritics are removed. Character combinations are mapped to the base characters (*uͦ*, *æ* become *uo, ae*, respectively). Abbreviations are spelt out (e.g., the ʾ-hook becomes *er*).
- Finally, the *norm*(alized) version is an artificial standard form, similar to the citation forms used in lexicons of MHG, such as Lexer (1872). The normalized form abstracts away completely from dialectal sound (grapheme) variance. It has been semi-automatically generated by a tool developed by the project "Mittelhochdeutsche Grammatik". The tool exploits lemma and morphological information in combination with symbolic rules that encode linguistic knowledge about historical dialects (Klein, 2001). We use a simplified ASCII version of the normalized form, with modifications similar

[4] <ʾ> can also stand for *re, r*, and rarely for *ri, ir*. We replace it unambiguously by *er*, which seems to be the most frequent case.

Texts	Tokens	Types		
		strict	*simple*	*norm*
51 total	211,000	40,500	34,500	20,500
		.19	.16	.10
27 MG	91,000	22,000	19,000	13,000
		.24	.21	.14
20 UG	67,000	15,000	13,500	8,500
		.22	.20	.13
4 mixed	53,000			

Table 1: Number of tokens and types in the Middle High German corpus. Below each type figure, the type-token ratio is given.

to the ones of the simple transcription version.

Table 1 displays some statistics of the current state of the corpus. The first column shows that there are currently 51 texts in total, with a total of around 211,000 tokens. The shortest text contains only 51 tokens, the longest one 25,000 tokens. 27 texts are from MG dialects and 20 from UG dialects. 4 texts are classified as "mixed", because they show mixed dialectal features, or are composed of fragments of different dialects.

As Table 1 shows, the numbers of types are somewhat reduced if strict (diplomatic) word forms are mapped to simple forms. Comparing strict and normalized types, the numbers are roughly cut in half. This benefits current taggers, as it reduces the problem of data sparseness to some extent. The question is, however, how reliably the normalized form can be generated automatically. The current tool requires a considerable amount of manual intervention during the analyses of lemma and morphology.

MG texts seem more diverse than UG texts: Despite the fact that the MG subcorpus is larger than the UG subcorpus, it has a higher type/token ratio.

The texts are further annotated with POS tags. The original POS tagset comprises more than 100 tags and encodes very fine-grained information. For instance, there are 17 different tags for verbs, whose main purpose is to indicate the inflection class that the verb belongs to. For the experiments described in this paper, these POS tags were mapped automatically to a modified version of the STTS tagset, the de-facto standard tagset for modern German corpora (Schiller et al. (1999); see Fig. 1).[5]

3 The Experiments

For the experiments, we performed a 10-fold cross-validation. The split was done in blocks of 10 sentences (or "units" of a fixed number of words, if no punctuation marks were available[6]). Within each block, one sentence was randomly extracted and held out for the evaluation.

For the analysis, we used the TreeTagger, since it takes suffix information into account. Thanks to this property, the TreeTagger can profit from units smaller than words, which seems favorable for data with high variance in spelling.

In our experiments, we modified two parameters during training: (i) word forms: *strict, simple, norm*; (ii) dialects: *all, MG, UG* (i.e., training data consists of the entire corpus, or the MG subcorpus, or the UG subcorpus). For instance, in one setting the tagger is trained on the strict forms of MG data.

For the evaluation, we introduced a further parameter: (iii) tagger: *specific, general, incorrect.* In the specific setting, the trained tagger is applied to the "correct", specific data (e.g., the tagger trained on strict-MG data is evaluated on strict-MG data). In the general setting, the tagger trained on the entire corpus is applied to some subcorpus. Finally, in the "incorrect" setting, the tagger trained on MG data is evaluated on UG data, and vice versa.

The first evaluation setting is straightforward. Setting two gives an impression of which performance we can expect if we apply a tagger that has been trained on a larger data set, which, however, consists of heterogeneous dialect texts. Setting three shows the extent to which performance can de-

[5] The STTS modifications are:
(i) An underspecified tag for the demonstrative or relative pronoun *der* has been introduced. The distinction between both types of pronouns can be made rather easily for modern German texts: relative pronouns induce subordinate word order, whereas demonstrative pronouns do not. In MHG, the position of the verb, which marks subordinate word order, was not as fixed as nowadays. Hence, this property should not be used as a criterion.
(ii) General tags PW, PI, and KO are used rather than PWS/PWAT, or PIS/PIAT, or KON/KOUS etc., because the original tagset does not allow to reconstruct the distinction.
(iii) All adjectives are tagged with the underspecified tag ADJ. Predicative adjectives can be inflected in MHG and, thus, a mapping to ADJA/ADJD is not easily definable.
(iv) Finally, the suffix _LAT subtype was introduced to mark Latin words and their POS tags (e.g. V_LAT for Latin verbs). These occur quite frequently in historical texts. Our corpus contains a total of 5,500 Latin words (= 2.6% of all tokens). In the MG texts, 5.3% of the tokens are Latin, whereas in the UG texts, only 0.9% are Latin.

[6] Punctuation marks in historical texts do not necessarily mark sentence or phrase boundaries. Nevertheless, they probably can serve as indicators of unit boundaries at least as well as randomly-picked boundary positions.

Dialect	Tagger	Word Forms		
		strict	*simple*	*norm*
MG	*specific*	86.62 ± 0.63	87.65 ± 0.59	91.43 ± 0.39
	general	86.92 ± 0.64	87.69 ± 0.58	91.66 ± 0.47
	incorrect	65.48 ± 0.73	71.20 ± 0.56	81.59 ± 0.44
	unknowns	59.71 ± 1.84	62.26 ± 2.18	68.14 ± 1.34
UG	*specific*	89.16 ± 0.75	89.58 ± 0.72	**92.91 ± 0.29**
	general	88.88 ± 0.68	89.45 ± 0.59	92.83 ± 0.39
	incorrect	77.81 ± 0.80	79.76 ± 0.57	89.43 ± 0.49
	unknowns	62.77 ± 1.77	64.81 ± 2.62	70.46 ± 2.21

Table 2: Results of a 18 test runs, based on different types of word forms, dialect subcorpora, and taggers, and 6 evaluations of unknown tokens. For each scenario, mean and standard deviation of per-word accuracy across the 10 folds are given (all values are percentages).

grade in a kind of worst case scenario. In addition, settings three allows us to compare the impact of the normalization step: Since normalization is supposed to level out dialectal differences, we expect less deterioration of performance with norm forms than with strict or simple forms.

The results of the different scenarios are summarized in Table 2. For each scenario, mean and standard deviation of per-word accuracy across the 10 folds are given.

The table shows that taggers achieve higher scores with UG data than MG data, in all scenarios. This result is somewhat unexpected since the size of the UG subcorpus is only 75% of that of the MG subcorpus. However, as we have seen, MG data is more diverse and has a higher type/token ratio.

With respect to the different types of word forms, tagging with normalized forms turns out best, as expected. The differences between strict and simple forms are surprisingly small, given the fact that the size of the "simpler vocabulary" is only around 85% of the "strict vocabulary".[7] The wide difference between simple and normalized forms reflects the fact that the level of standardization as introduced by the normalization step concerns not just minor features such as accents or ligatures but also inflectional endings and sound changes.

Comparing the three types of taggers, the table clearly shows that the specific and general taggers perform considerably better than the incorrect ones. As expected, the differences between the taggers are less pronounced with normalized word forms. Interestingly, the specific and general tagger variants do *not* differ significantly in most of the scenarios, despite the fact that the general taggers have been trained on a larger data set.[8]

Finally we evaluated the performance of the specific taggers on unknown words. The results show that performance degrades considerably, by 22.5 (with the UG-norm tagger) up to 26.9 percentage points (with the MG-strict tagger).

4 Summary and Outlook

We presented a set of experiments in POS tagging of historical data. The aim of this enterprise is to evaluate how well a state-of-the-art tagger, such as the TreeTagger, performs in different kinds of scenarios. The results cannot directly compared to results from modern German, though: Our corpora are rather small; historical data is considerably more diverse than modern data; and we used a modified version of the STTS.

As future steps, we will perform a detailed error analysis: Which tags are especially hard to learn, which tags are difficult to distinguish? Can certain errors be traced back to dialectal properties of the language? Is there an impact of time of origin of a manuscript?

To reduce variance of the data, without requiring complete normalization of all words, we plan to investigate a hybrid approach, by evaluating whether it is helpful to normalize function words only and keep content words unmodified. Since function words are closed classes, it might be possible to successfully normalize these words automatically, without manual intervention.

[7] Maybe this outcome can be attributed to the TreeTagger, which possibly performs similar simplifications internally. All word form results differ significantly from each other, though.

[8] The general taggers perform significantly better than the corresponding specific taggers when they are evaluated on MG-norm and MG-strict data (paired t-test; MG-norm data: t=4.48, df=9, p<.01; MG-strict data: t=4.20, df=9, p<.01).

References

Thomas Klein. 2001. Vom lemmatisierten Index zur Grammatik. In Stephan Moser, Peter Stahl, Werner Wegstein, and Norbert Richard Wolf, editors, *Maschinelle Verarbeitung altdeutscher Texte V. Beiträge zum Fünften Internationalen Symposion Würzburg 4.-6. März 1997*, pages 83–103. Tübingen: Niemeyer.

Anthony Kroch and Ann Taylor. 2000. Penn-Helsinki parsed corpus of Middle English. Second edition, `http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/`.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. 2004. Penn-Helsinki parsed corpus of Early Modern English. `http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/`.

Matthias Lexer. 1872. *Mittelhochdeutsches Handwörterbuch*. Leipzig. 3 Volumes 1872–1878. Reprint: Hirzel, Stuttgart 1992.

Thomas Pilz, Wolfram Luther, Ulrich Ammon, and Norbert Fuhr. 2006. Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*, 21:179–86.

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 *Proceedings of Corpus Linguistics 2007*, University of Birmingham, UK.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, University of Stuttgart and University of Tübingen.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.