Stefanie Dipper

Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison

This paper deals with morphological and part-of-speech tagging applied to manuscripts written in Middle High German. I 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. Normalization improves accuracies by 3.56–7.10 percentage points, resulting in accuracies of $> 79\%$ for morphological tagging, and $> 91\%$ for part-of-speech tagging. Comparing Middle with New High German data of similar size, the evaluation shows that part-of-speech tagging, but not morphological tagging, is clearly easier with modern data.

1 Introduction[1]

This paper deals with automatic analysis of historical language data, namely morphological and 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 and, 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 statistical analysis tools, whose quality usually depend on the availability of large training samples. Automatic 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 with POS tags in a bootstrapping approach, which involves successive cycles of manual annotation, training, automatic tagging, followed by manual corrections, etc. Rayson et al. (2007) and Pilz et al. (2006) map historical word forms to the corresponding modern word forms, and analyze these by state-of-the-art POS taggers. The mappings make use of the Soundex algorithm,

[1] I 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.

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.

Morphological tagging has received far less attention than POS tagging, presumably because English, which is the most researched language in computational linguistics, does not have rich morphology, and, furthermore, a considerable amount of (overtly marked) morphological information is in fact recorded by common English POS tagsets, e.g. for nouns: singular vs. plural form, for verbs: uninflected base form vs. third-singular present tense vs. past tense vs. participle, etc. Similar coarse-grained distinctions have been transferred to languages with rich(er) morphology, such as German. For instance, in the de-facto standard tagset for modern German corpora, the STTS (Schiller et al., 1999), all finite verb forms receive the tag VVFIN ("full verb, finite"), infinitives the tag VVINF ("full verb, infinitive"), etc. However, in contrast to English, the tag VVFIN covers up to five differently-inflected verb forms; similarly, the tag NN ("common noun"[2]) also corresponds to up to five different forms. Hence, full morphological tagging, which would differentiate between the different forms, could provide valuable information in languages with rather free word order: morphological information can help in determining constituents and grammatical functions. POS and morphological tagging thus represents important preprocessing steps, e.g., for treebanking or natural language processing of such languages.

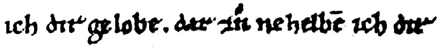
This paper reports on experiments in applying a state-of-the-art tagger, the TreeTagger (Schmid, 1994), to a corpus of texts from Middle High German (MHG).[3] The tagger is used for both morphological and POS tagging. My approach is similar to the one by Kroch et al. in that I train and apply the tagger to historical rather than modern word forms. The tagging experiments make use of a balanced MHG corpus that is created and annotated in the context of two projects, the projects "Mittelhochdeutsche Grammatik" and "Referenzkorpus Mittelhochdeutsch".[4] The corpus has been semi-automatically annotated with morphology, POS tags, lemma, and a normalized word form, which represents a virtual historical standardized form. The corpus is not annotated with modern word forms.

I present the results of a set of experiments that involve different types of tokens (original and normalized versions) and dialect-specific subcorpora. Sec. 2 gives detailed information about the corpus and its annotations, Sec. 3 addresses the tagging experiments and results. In many places, I contrast the historical data with a modern corpus, the TIGER corpus (Brants et al., 2004). Sec. 4 presents a summary.

[2] Tags for nouns in German tagsets are usually unspecified for number.

[3] In a recent evaluation of part-of-speech taggers on German web data, Giesbrecht and Evert (2009) found that the Stanford tagger (Toutanova et al., 2003) performed best (97.63%) while the TreeTagger only achieved an accuracy of 96.89%. On the other hand, training the taggers took 10 seconds (TreeTagger) vs. 5.5 hours (Stanford). Another important advantage of the TreeTagger is the fact that its model can be inspected and easily interpreted (the options "-print-prob-tree" and "-print-suffix-tree" print out the decision tree for ngrams and the suffix lexicon, respectively). Moreover, training the TreeTagger is straightforward and does not require any specific preprocessing, in contrast, e.g., to the RFTagger (Schmid and Laws, 2008), which presupposes the definition of a finite-state automaton for the tag labels.

[4] http://www.mittelhochdeutsche-grammatik.de and http://www.linguistics.rub.de/mhd/.



Dipl	ich	dir	gelobe	.	dar	zuͦ	ne	helbē	ich	dir
Norm	ich	dir	gelobe	.	dar	zuo	ne	hilfen	ich	dir
Lemma	ich	dû	ge-loben		dâr	zuo	ne	hëlfen	ich	dû
Morph	*.Nom.Sg	*.Dat.Sg	1.Sg.Pres.*	–	–	–	–	1.Sg.Pres.Ind	*.Nom.Sg	*.Dat.Sg
Pos	PPER	PPER	VVFIN	$.	ADV	ADV	NEG	VVFIN	PPER	PPER
Gloss	I	you	promise		there	to	not	help	I	you

Figure 1: A line from Eilhart's *Tristrant* (Magdeburg fragment), along with a diplomatic transcription, normalized word forms, and linguistic annotations. The complete sentence is: *vil ernirsthafte ich dir gelobe. dar zuo ne helben ich dir niet* 'Very seriously I promise you: I do not help you with this'.

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 optimally as possible. The texts distribute in time, i.e. over the relevant centuries, and in space, coming from a variety of Central German (CG) and Upper German (UG) dialects. CG dialects were spoken in the central part of Germany; examples are Franconian or Thuringian. UG dialects were (and are still) spoken in Southern Germany, Switzerland, and Austria, e.g. Swabian, Alemannic, or Bavarian.

The corpus provides two different versions of "word forms": the diplomatic transcription and a normalized form. Figure 1 presents an example fragment encoded in the different versions.[5] Below the lines with the word forms, linguistic annotations are displayed: lemma, morphology, parts of speech (POS).

Lines DIPL and NORM 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, superscribed letters (e.g. uͦ), variant letter forms (e.g. short vs. long s: <s> vs. <ſ>), or abbreviations (e.g., the superscribed "nasal bar" <¯> substitutes *n*).[6]

[5] The manuscript screenshot has been taken from `http://www.hs-augsburg.de/~harsch/germanica/Chronologie/12Jh/Eilhart/eil_tmma.html`

[6] Internally, I use an isomorphic ASCII-encoded representation of the diplomatic transcription. 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*.

Corpus Dialect (#Texts)		Tokens	Types and TTR *dipl*	 *norm*
total	(51)	211,000	40,500 .19	20,500 .10
CG	(27)	91,000	22,000 .24	13,000 .14
UG	(20)	67,000	15,000 .22	8,500 .13
mixed	(4)	53,000		

Corpus	Tokens	Types and TTR
TIGER	1,000,000	81,000 .09
	210,000	30,000 .14
	90,000	16,000 .18

Table 1: Number of tokens and types in the Middle High German corpus (left) and in differently-sized subcorpora of the TIGER corpus (right). Below each type figure, the type-token ratio (TTR) is given.

The *normalized* version is an artificial standard form, similar to the citation forms used in lexicons of MHG, such as Lexer (1872).[7] The normalized form abstracts away completely from dialectal sound (grapheme) variance. It has been semi-automatically generated by a tool developed by Thomas Klein (Klein, 2001) within the project "Mittelhochdeutsche Grammatik". The tool exploits lemma and morphological information in combination with symbolic rules that encode linguistic knowledge about historical dialects. The user has to provide information about the dialect of the text, and to correct intermediate results interactively. No information about overall accuracy or inter-annotator agreement is available.

Table 1 displays some statistics of the current state of the corpus (left table). 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 CG 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. Due to their nature, the mixed texts have been excluded from detailed consideration.

The table shows that the numbers of types are considerably reduced if diplomatic word forms are mapped to normalized forms. 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 analysis.

[7] Internally, I use a simplified ASCII version of the normalized form, with the following modifications: Umlaut has been replaced by the corresponding voyel + e (e.g. "ä" becomes "ae"); other accents or diacritics have been removed.

CG texts seem more diverse than UG texts: despite the fact that the CG subcorpus is larger than the UG subcorpus, it has a higher type/token ratio (TTR). Usually longer texts tend to have lower TTR values. This is shown by the right table of Table 1: The entire TIGER corpus (1,000,000 tokens) has a TTR of .09, i.e., there are 11.1 corpus instances of each word (type) on average. Taking into account only the first 210,000 tokens of the TIGER corpus, TTR goes up to .14; this corresponds to 7.1 instances of each word on average. The TTR of the 90,000 TIGER subcorpus, which is comparable in size with the CG subcorpus, shows that New High German (NHG, i.e. newspaper texts from the 1990s) has a more diverse vocabulary than the MHG texts.

Judging from these figures, one could predict the following outcomes:[8]

1. Normalized vs. diplomatic: Tagging **normalized** data should be easier
2. CG vs. UG vs. NHG data:
   a) Tagging **CG** should be easier that UG, because more training data is available
   b) Alternatively: tagging **UG** is easier than CG, because it is less diverse (has a lower TTR)
   c) Tagging (equally-sized subsets of) **MHG** should be easier than NHG, because it has lower TTRs

Line MORPH In addition to normalized word forms, the texts have also been annotated with morphological and part-of-spech (POS) tags, by the tool by Klein (2001). The original morphological tagset consists of around 430 tags. The large number of tags is partly due to the fact that inherent gender of nouns was not yet as fixed as it is nowadays. That is, many nouns could be used, e.g., with masculine or feminine articles (or with all three genders). In all cases where the context does not allow for gender disambiguation, ambiguous tags have been annotated, as in Ex. (1). “MascFem.Nom.Pl” means nominativ plural, masculine or feminine. “*” means that a feature is entirely underspecified, such as gender with the plural pronoun *sie* ‘them’, which is therefore tagged as “*.Acc.Pl”.

(1)	daz	si	slangen	bizzen
	—	*.Acc.Pl	MascFem.Nom.Pl	3.Pl.Past.*
	that	them	snakes	bit

‘that snakes bit them’

Moreover, properties such as postnominal position, e.g., of adjectives or possessive determiners, or morphological unmarkedness, have also been recorded by the original tagset. For the experiments described in this paper, these morphology tags were mapped automatically to a slightly modified version of the STTS morphological tagset. (If the value of a specific slot could not be determined automatically, it was also filled by “*”.)

[8] Of course, the outcomes also depend on properties of the tagsets, see below.

Corpus	Morphology			Part of Speech		
	#Tags	Tags/Word	$\tilde{x}$ (max)	#Tags	Tags/Word	$\tilde{x}$ (max)
CG *norm*	245	Ø1.40 ± 1.16	1 (23)	44	Ø1.10 ± 0.37	1 (7)
UG *norm*	219	Ø1.46 ± 1.28	1 (33)	41	Ø1.10 ± 0.35	1 (6)
TIGER						
1,000 K	270	Ø1.48 ± 1.22	1 (40)	54	Ø1.05 ± 0.25	1 (7)
210 K	230	Ø1.37 ± 0.97	1 (26)	53	Ø1.05 ± 0.23	1 (6)
90 K	205	Ø1.32 ± 0.86	1 (18)	51	Ø1.04 ± 0.21	1 (6)

Table 2: Sizes of the tagsets and average number of tags per word (with standard deviation), as occurring in the normalized training data, along with the median ($\tilde{x}$) and maximum.

Line POS The original POS tagset comprises more than 100 tags and, similarly to the morphological tagset, 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 POS tagset (for a description of the modifications, see Dipper (2010, Fn.5)).

Table 2 presents relevant statistical information about the resulting STTS-based tagsets. One can see that the sizes of the tagsets are similar with CG, UG, and NHG data. Morphological tagsets are 4–5.5 times larger than POS tagsets. Historical data in general seems more ambiguous than modern data, on average. The figures have to be interpreted with care, though, because the tagsets cannot be directly compared: there is no isomorphic mapping between the information encoded by the original MHG tagsets and the STTS tagsets, and underspecified tags have to be used in the MHG data rather often.

The figures also confirm that the sizes of the corpora are rather small: numbers calculated from the TIGER subcorpora show that adding more data increases the number of tags occurring in the data, especially in the case of morphological tags. That is, even in the complete TIGER corpus, not all available (morphological) tags do occur at least once.[9]

Despite these caveats, we could add the following predictions, based on the figures in Table 2:

3. Morphology vs. POS:
   Tagging of **POS** information should be easier (due to a lower ambiguity rate)

[9] As defined in the header of the TIGER corpus, the total number of morphological STTS tags is 585. Presumably, however, a good amount of them are theoretically possible tags but without any actual instance in the language.

4. CG vs. UG vs. NHG data:
   a) Tagging **NHG** data should be easier (due to a lower ambiguity rate) — this is contrary to the expectation formulated above (see Prediction 2c).
   b) Results for CG and UG should be comparable (almost identical average of ambiguity rates). — The situation here is similar to above: no clear advantage emerges (cf. Predictions 2a and 2b).
   c) However, UG has a higher maximum with ambiguous morphology tags, CG with ambiguous POS tags. Hence, **CG** could perform better with morphological tagging than UG, and **UG** could perform better with POS tagging than CG.

3 Experiments and Results

For the experiments with the historical data, I 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[10]). Within each block, one sentence was randomly extracted and held out for the evaluation.

For the analysis, I used the TreeTagger. It takes suffix information into account so that it can profit from units smaller than words. This seems favorable for data with high variance in spelling. Moreover, the TreeTagger allows the user to inspect the ngram and suffix models acquired during training (see Fn. 3).

In the experiments, I varied two parameters concerning the input data ("dialect, word forms") and one parameter concerning training ("tagger"):

1. Dialect: *CG, UG*

2. Word forms: *dipl, norm*
   For instance, in one setting input data consists of normalized data from Central German (CG-*norm*).

3. Tagger: *gen*(eric), *spec*(ific). In the generic setting, the tagger is trained on the entire corpus (210,000 tokens) and then evaluated on the CG and UG subcorpora. In the specific setting, the tagger is trained and evaluated on the subcorpora only (e.g., the tagger is trained and evaluated on CG-*norm* data). This allows us to evaluate whether a larger set of training data is favorable to a set that is smaller but more homogeneous.

Furthermore, as I have discussed in Sec. 1, POS tags already encode a considerable amount of morphological information. Hence, to improve accuracy with morphological tagging, I also fed the tagger with preprocessed data, which contained POS annotations, so that the morphological tagger could profit from that information.

[10] 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.

Since I wanted to use the TreeTagger in all experiments, there were two options to integrate POS information in the input data. First, morphological and POS tags can be presented in turn, as shown in (ii) below. Second, POS tags could be appended as suffixes to wordforms, as in (iii). With the first option, the TreeTagger would make use of POS information in its ngram model; with the second option, the suffix lexicon would record POS-morphology dependencies. (i)–(iii) show example input for all three scenarios, for the sequence *werde disemo* 'would this'.

(i) *No use* of POS; input example:

```
werde    3.Sg.Pres.Subj
disemo   Neut.Dat.Sg
```

(ii) *Successive pairs* of <word, morph><word, POS>:
(or vice versa: <word, POS><word, morph>):

```
werde    3.Sg.Pres.Subj
werde    VAFIN
disemo   Neut.Dat.Sg
disemo   PD
```

(iii) *Merged pairs* of <word.POS, morph>:

```
werde.VAFIN    3.Sg.Pres.Subj
disemo.PD      Neut.Dat.Sg
```

The task based on successive pairs seems harder than the task with merged pairs: Successive pairs involve learning POS and morphology assignments simultaneously. With merged pairs, in contrast, the POS tags are given (as part of the word forms). However, to make the scenario realistic, the POS tags of the evaluation data have been assigned automatically and, hence, are incorrect to a certain extent. To assess the impact of incorrect POS tags, I repeated the evaluation of Scenario (iii) with gold POS annotations, which gives us an upper bound of the approach.

The results of the different scenarios are summarized in Table 3. For each scenario, mean and standard deviation of per-word accuracy across the 10 folds are given.[11] I now check the predictions from Sec. 2 against the figures in Table 3.

Prediction 1: Tagging normalized data should be easier Tagging with normalized word forms turns out better, as expected. This holds for both morphological and POS tagging.[12] Improvements are more pronounced with CG data (4.74–7.10 percentage points) than with UG data (3.56–5.36). There is no obvious explanation for this

[11] Evaluation of Scenarios (ii) and (iii) only considers morphological tags. Reordering the pairs as POS > morph resulted in slightly lower accuracy (< 1.6 percentage points). A more detailed evaluation of tagging POS can be found in Dipper (2010).

[12] Normalization resulted in a highly significant increase of accuracy in all scenarios (paired t-test; p<.001).

Morphology Scenario	Dialect	Tagger	Word Forms	
			diplomatic	*normalized*
(i) No use	CG	*gen*	73.91 ± 0.51	***79.70*** ± 0.36
		spec	72.64 ± 0.54	78.43 ± 0.53
	UG	*gen*	73.85 ± 1.16	78.28 ± 1.71
		spec	73.23 ± 1.02	78.15 ± 1.28
	TIGER 1,000 K		—	79.08
	210 K	≈ *gen*	—	76.95
	90 K	≈ *spec*	—	75.71
(ii) Successive pairs (morph > POS)	CG	*gen*	74.23 ± 0.51	**80.84** ± 0.55
		spec	72.37 ± 0.51	79.47 ± 0.50
	UG	*gen*	74.17 ± 1.10	79.11 ± 1.51
		spec	73.27 ± 0.96	78.63 ± 1.30
(iii) Merged pairs	CG	*gen*	74.39 ± 0.50	***79.81*** ± 0.42
		spec	72.86 ± 0.36	78.48 ± 0.53
	UG	*gen*	74.07 ± 0.88	77.63 ± 1.99
		spec	73.14 ± 0.85	77.02 ± 1.69
(iv) Gold POS (with (iii))	CG	*gen*	77.14 ± 0.47	*82.19* ± 0.39
		spec	75.54 ± 0.40	80.80 ± 0.49
	UG	*gen*	76.79 ± 0.87	80.83 ± 1.56
		spec	75.79 ± 0.86	80.26 ± 1.22

Part of Speech	Dialect	Tagger	Word Forms	
			diplomatic	*normalized*
	CG	*gen*	86.92 ± 0.64	91.66 ± 0.47
		spec	86.62 ± 0.63	91.43 ± 0.39
	UG	*gen*	88.88 ± 0.68	92.83 ± 0.39
		spec	89.16 ± 0.75	**92.91** ± 0.29
	TIGER 1,000 K		—	95.81
	210 K	≈ *gen*	—	95.67
	90 K	≈ *spec*	—	94.39

Table 3: Results of different test runs for morphological tagging (table on top) and POS tagging (table at the bottom), based on different types of word forms, dialect subcorpora, and taggers. For each scenario, mean and standard deviation of per-word accuracy across the 10 folds are given (all values are percentages). The overall best results for morphological and POS tagging of MHG data are indicated in bold, best results for other scenarios in bold italics. Results of Scenario (iv) represent an upper bound. Selected results from simple training (no cross-validation/standard deviation) on NHG (TIGER) are added for comparison. Training data of 210 K corresponds to the training data of the generic tagger, 90-K-training data corresponds to the data of the CG-specific tagger.

difference — with both dialect subcorpora, the type-token ratios are almost cut in half with normalized data.

Comparing the two types of taggers, generic vs. specific, the tables show that the generic taggers almost always perform better than the specific ones (the exception is POS tagging of UG). This seems to indicate that enlarging the training set is favorable even if the input becomes more heterogeneous. However, the differences in accuracy are rather small in general, and not significant in some of the scenarios.[13]

Predictions 2a / b: CG data / UG data is easier to tag Judging from the morphological top results, performance on CG data is slightly superior to performance on UG data (Prediction 2a). However, most of the differences are not significant.[14] On the other hand, UG data yields the best result with POS tagging (Prediction 2b; highly significant differences). Maybe this "contradiction" can be attributed to the fact that the morphological ambiguity rate is more favorable for CG data (lower mean and smaller standard deviation and maximum than UG data), while the opposite is true of the POS ambiguity rate.

Predictions 2c / 4: Tagging MHG / NHG should be easier Looking at the morphology table, we see that tagging of MHG data indeed outperforms tagging of NHG data (thus confirming Prediction 2c). Turning to the morphology table, the picture is, again, reversed (thus confirming Prediction 4): NHG tagging is well above MHG tagging. When the training size is reduced, accuracy of NHG degrades to a certain extent, but clearly remains superior. As above, the discrepancy can be traced back to ambiguity rates, which favour morphology tagging of MHG data, and POS tagging of NHG data.

Prediction 3: POS tagging should be easier Prediction 3 is clearly borne out. The gap between morphological and POS tagging is more than 10 percentage points:

– Morph (Scenarios (i)–(iii)):	> 79% (CG-*norm*), > 77% (UG-*norm*)
– POS:	> 91% (CG-*norm*), > 92% (UG-*norm*)

Interestingly, Scenario (iii) is not superior to Scenario (i), which makes no use of POS tags at all. This seems to suggest that automatically-assigned POS tags could not improve morphological tagging. However, the results from Scenario (ii) show that some improvement can indeed be achieved.

[13] The differences between the generic taggers and the corresponding specific taggers are *not* significant when they are evaluated on data from UG-*norm* (morphology Scenario (i) and POS), and UG-*dipl* (POS) (paired t-test).

[14] The differences between CG and UG taggers *are* significant with the generic taggers applied to normalized data, in all scenarios (paired t-test; p<.01 to p<.05).

4 Summary

I presented a set of experiments in morphological and 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: The corpora are rather small; historical data is considerably more diverse than modern data; and I used modified versions of the STTS.

To summarize the main results from the set of experiments: Simple training on historical data results in satisfiable results of $> 91\%$ accuracy for POS tagging. In contrast, morphological tagging ($> 79\%$ accuracy) needs more sophisticated methods. For instance, the RFTagger (Schmid and Laws, 2008) is able to analyze and decompose complex morphological tags and, thus, to reduce the problem of data sparseness that arises especially with large, fine-grained tagsets (but see Fn. 3). Normalization increases accuracy by 3.56–7.10 percentage points.

The evaluations show that fully-automatic annotations (without subsequent manual corrections) currently only make sense with POS taggers, but not (yet) with morphological taggers. Assuming that automatic annotations would be checked manually, it is interesting to know how many correct tags are among the top *n* most probable tags. If most of the time, the correct tag is easy to select, in an efficient way, the current performance of the taggers might not be such a problem, after all.

I computed the ranks of all correct tags for a CG-*norm* sample, tagged with morphology, Scenarios (iii), and POS, see Table 4. The morphology table shows that in 87.1% of the cases, the correct tag is among the top-3 ranks (POS: 96.2%).[15] This means that it would probably speed up the annotation process if human annotators were presented the first three most probable tags to choose from.

As a next step, I want to evaluate the RFTagger for tagging of historical data. In addition, I plan to perform a detailed analysis with the goal of relating the tagging results to linguistic features of the different dialects.

References

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

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

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the German Web as Corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 27–35.

[15] I set the probability threshold to .1, i.e., all tags with a probability higher than 10% of the probability of the best tag are output. Scenario (ii) cannot be easily evaluated in this respect, because the probabilities are distributed over both morphological and POS tags.

Morphology (iii)		
Rank	# Word forms	
1	7467	79.6%
2	600	6.4%
3	99	1.1%
None	1122	12.0%

Part of Speech		
Rank	# Word forms	
1	8160	92.0%
2	370	4.2%
None	303	3.4%

Table 4: Ranks of the correct tags, which have been sorted according to their probabilities (left: morphology, Scenario (iii), right: POS). Absolute and relative frequencies are given (no cross-validation). Rank "None" shows the number of word forms whose actual tag is not among the automatically-proposed tags. Ranks with less than 1% instances are not displayed.

Klein, T. (2001). Vom lemmatisierten Index zur Grammatik. In Moser, S., Stahl, P., Wegstein, W., and Wolf, N. R., editors, *Maschinelle Verarbeitung altdeutscher Texte V. Beiträge zum Fünften Internationalen Symposion Würzburg 4.-6. März 1997*, pages 83–103. Tübingen: Niemeyer.

Kroch, A., Santorini, B., and Delfs, L. (2004). Penn-Helsinki parsed corpus of Early Modern English. `http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/`.

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

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

Pilz, T., Luther, W., Ammon, U., and Fuhr, N. (2006). Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*, 21:179–86.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (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.

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

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

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.