POS Tagging for Historical Texts with Sparse Training Data

Marcel Bollmann
Department of Linguistics, Ruhr University Bochum
bollmann@linguistics.rub.de

Abstract

This paper presents a method for part-of-speech tagging of historical data and evaluates it on texts from different corpora of historical German (15th–18th century). Spelling normalization is used to preprocess the texts before applying a POS tagger trained on modern German corpora. Using only 250 manually normalized tokens as training data, the tagging accuracy of a manuscript from the 15th century can be raised from 28.65% to 74.89%.

1 Introduction[1]

Part-of-speech (POS) tagging of modern language data is a well-explored field, commonly achieving accuracies around 97% (Brants, 2000; Schmid and Laws, 2008). For historical language varieties, the situation is worse, as specialized taggers are typically not available. As an example, a study by Scheible et al. (2011a) reports an average tagging accuracy of 69.6% for Early Modern German texts. However, with projects to create historical corpora being on the rise (Sánchez-Marco et al., 2010; Scheible et al., 2011b, are recent examples), the need for more accurate tagging methods on these types of data increases.

A common approach for historical texts is to use spelling normalization to map historical wordforms to modern ones (Baron and Rayson, 2008; Jurish, 2010). Manually normalized data was found to improve POS tagging accuracy for a variety of languages such as German, English, and Portuguese, with accuracies between 79% and 91% (Scheible et al., 2011a; Rayson et al., 2007; Hendrickx and Marquilhas, 2011).

[1]I would like to thank the anonymous reviewers for their helpful comments. The research reported here was supported by Deutsche Forschungsgemeinschaft (DFG), Grants DI 1558/4-1 and DI 1558/5-1.

This paper presents results for POS tagging of historical German from 1400 to 1770, classified here as Early New High German (ENHG), using automatic spelling normalization to preprocess the data for a POS tagger trained on modern German corpora. To train the normalization tool, short fragments of a few hundred tokens are used for each text. This approach allows for a better adaptation to the individual spelling characteristics of each text while requiring only small amounts of training data. Additionally, different ways to deal with typical obstacles for processing historical texts (e.g., inconsistent use of punctuation) are compared.

The structure of this paper is as follows. Sec. 2 presents the historical texts used for the evaluation. Sec. 3 describes the approach to normalization, while Sec. 4 discusses problems and results of POS tagging on normalized data. Sec. 5 presents related work, and Sec. 6 concludes.

2 Corpora

This study considers texts from two corpora of historical German: the Anselm corpus (Dipper and Schultz-Balluff, 2013) and the GerManC-GS corpus (Scheible et al., 2011b).

The Anselm corpus consists of more than 50 different versions of a medieval religious treatise written up in various German dialects. As the creation of gold-standard annotations for the corpus is still in progress, only two texts are used here: a manuscript in an Eastern Upper German dialect kept in Melk, Austria; and an Eastern Central German manuscript kept in Berlin. Both manuscripts are dated to the 15th century.

The GerManC-GS corpus aims to be a representative subcorpus of GerManC with additional gold-standard annotations. It contains texts from Early Modern German categorized by genre, region, and time period. For this study, the three texts of the genre “sermon” are used. They are

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 11–18,
Sofia, Bulgaria, August 8-9, 2013.

Corpus	Date	Name	Tokens
Anselm	15c	Berlin	5,399
	15c	Melk	4,783
GerManC-GS	1677	LeichSermon	2,585
	1730	JubelFeste	2,523
	1770	Gottesdienst	2,292

Table 1: Texts used for the evaluation

dated from 1677 to 1770, which makes them considerably newer than the Anselm texts. Table 1 gives an overview of all texts used here.

All texts are manually annotated with normalizations and POS tags. In the normalization layer, tokens are mapped to modern German equivalents. The normalization schemes are not identical, but roughly comparable for both GerManC-GS and Anselm (see Scheible et al. (2011b) and Bollmann et al. (2012) for details). In both corpora, POS tagging follows the STTS tagset (Schiller et al., 1999) without morphological information, though some additional tags were introduced in GerManC-GS. For our evaluation, they are mapped back to standard STTS tags; this mapping only affects 80 tokens from all three texts.

Additionally, both corpora are annotated with modern punctuation and sentence boundaries; however, while modern punctuation is a separate annotation layer in Anselm, there is always a 1:1 correspondence between historical and modern (i.e., normalized) punctuation marks in GerManC-GS.

Finally, both corpora preserve many spelling characteristics of the original manuscripts, e.g., superposition of characters such as *å*, or abbreviation marks such as the nasal bar (as in *v̄*). Before any further processing, all wordforms are simplified to plain alphabetic characters; e.g., *å* is mapped to *uo*. For some abbreviation marks in the Anselm corpus, there is no clear "best" simplification: the nasal bar is a prime example here, which should be simplified most appropriately to *e*, *(e)n*, *(e)m*, or nothing, either before or after the letter on which it is placed, depending on context. In these cases, manually defined heuristics were used to guess the most appropriate mapping. As capitalization is not used consistently in the texts, all letters were additionally lowercased.

3 Normalization

Spelling normalization is performed using the Norma tool (Bollmann, 2012). It implements a chain of normalization methods—to the effect that methods further down the chain are only called if previous ones failed to produce a result—in the following order: (1) wordlist mapping; (2) rule-based normalization; and (3) weighted Levenshtein distance.

Wordlist mapping considers simple 1:1 mappings of historical wordforms to modern ones (e.g., *vnd* → *und* "and"), while rule-based normalization applies context-sensitive character rewrite rules (e.g., transform *v* to *u* between a word boundary and *n*) to an input string from left to right. Weighted Levenshtein distance assigns individual weights to character replacements (e.g., *v* → *u*), and performs normalization by retrieving the wordform from a modern lexicon which can be derived from the historical input wordform with the lowest cost, i.e., using a sequence of edit operations with the lowest sum of weights.

3.1 Normalization procedure

All normalization algorithms described above require some kind of parametrization to work (i.e., a wordlist; rewrite rules; Levenshtein weights). These parametrizations are neither hard-coded nor manually defined, but are derived automatically by the Norma tool from a set of manually normalized training data. For this purpose, short samples from the text to be normalized are used; i.e., training set and evaluation set are always disjoint parts of the same text. The reasons for choosing this approach lie in the individual spelling characteristics of the texts—the following examples show excerpts from Berlin, Melk, and LeichSermon, respectively, along with their (gold-standard) normalizations:

(1) *dyn lybes kynt*
dein liebes kind
"your dear child"

(2) *mein liebs chind*
mein liebes kind
"my dear child"

(3) *eins ihrer andern kinder*
eins ihrer anderen kinder
"one of their other children"

Text	Baseline	Normalizations			
		100	250	500	1,000
Berlin	23.05%	68.99%	75.02%	79.14%	81.83%
Melk	39.32%	69.10%	74.39%	75.74%	77.98%
LeichSermon	72.71%	77.96%	80.51%	82.85%	87.23%
JubelFeste	79.47%	88.50%	89.98%	91.87%	93.13%
Gottesdienst	83.41%	93.77%	95.24%	95.27%	95.56%

Table 2: Normalization accuracy after training on *n* tokens and evaluating on 1,000 tokens (average of 10 random training and evaluation sets), compared to the "baseline" score of the full text without any normalization

The first two examples, while both dated to the 15th century, show quite different spellings of the modern *Kind* "child": Ex. (1) shows the frequent use of *y* for modern *ei* or *i(e)*, while Ex. (2) demonstrates the frequent spelling *ch* for *k*. These differences are likely a cause of the different dialectal regions from which the manuscripts originate, but could also be attributed, at least in parts, to individual preferences by the manuscripts' writers. The LeichSermon text from 1677 in Ex. (3), on the other hand, already has the modern German spelling *Kind*.

Given this range of spelling variations, it seems implausible to achieve good normalization results using the same parametrization for each of the texts. Furthermore, for the older manuscripts showing more variation such as in Ex. (1), it is unclear what other training data could be used. The full GerManC-GS consists of texts from 1650 to 1800, while Jurish (2010) uses a corpus of German texts from 1780 to 1880; these texts are all considerably newer, consequently having less spelling variation than the Anselm texts. This lack of appropriate training data applies similarly to all kinds of less-resourced language varieties.

Therefore, while this approach requires slightly more effort for manually normalizing parts of the texts beforehand, it does not depend on the availability of a large training corpus or a specialized tool for the language variety to be processed.

3.2 Evaluation

Normalization is evaluated separately for each text, using a part of that text for training and evaluating on a different part of the same text. To address the question of how much training data is needed, evaluation is performed with different sizes of the training set in a range between 100 and 1,000 tokens. The evaluation set is kept at a fixed size of 1,000 tokens. Normalization accuracy is calculated by taking the average of 10 trials with randomly drawn training and evaluation sets. The results of this evaluation are shown in Table 2.

The baseline score for a text is defined as the percentage of matching tokens between the unmodified, historical text and its gold-standard normalization. There is a clear difference between the Anselm texts, with scores of 23% and 39%, and the GerManC-GS texts, which range from 72% to 83%. This shows that spelling variation affects significantly more wordforms in the Anselm texts. The age of a text is likely to be the main factor for this, as even within the group of GerManC-GS texts, a clear tendency for newer texts to have higher baseline scores can be observed.

Spelling normalization with the Norma tool shows rather positive results even for small training samples: with only 100 tokens used for training, it achieves a normalization accuracy of 69% for the Anselm texts, and raises the score for the GerManC-GS texts by 5–10 percentage points. Using 250 tokens results in another noticeable increase in accuracy, although the relative gain from increasing the training size even further attenuates after this point.

4 Part-of-speech tagging

While spelling normalization can be useful in itself (e.g., for search queries in the corpus), our main focus is on its usefulness for further processing of the data such as part-of-speech tagging. The results presented here were achieved using the RFTagger (Schmid and Laws, 2008) with an increased context size of 10, which we found to perform best on average on our data.

Text	OrigP	ModP	NoP
Berlin	85.78%	87.29%	87.07%
Melk	85.21%	87.76%	87.74%
LeichSermon	81.22%	80.59%	81.04%
JubelFeste	90.41%	90.41%	90.03%
Gottesdienst	93.24%	93.24%	92.27%

Table 3: Tagging accuracy on the gold-standard normalizations (OrigP = original punctuation, ModP = modern punctuation, NoP = no punctuation)

Original	96.85%
Lowercased	96.50%
No punctuation and SB	96.22%
Lowercased + no punctuation and SB	95.74%

Table 4: Tagging accuracy on the combined TIGER/Tüba corpus, using 10-fold CV, evaluated with and without capitalization, punctuation, and sentence boundaries (SB)

4.1 Impact of punctuation

Normalization tries to handle the problem of spelling inconsistencies found in historical language data. However, this is not the only challenge for processing the data with modern POS taggers. There is often no consistent capitalization, which can normally be used as a clue to detect nouns in modern German. This has already led to all wordforms being lowercased for the normalization process. Additionally, punctuation marks are also often used inconsistently or are missing completely: e.g., the Melk manuscript mostly uses virgules (visually resembling a modern slash '/') where modern German would use a full stop, but this is far from a definite rule, and large parts of the Anselm texts feature no punctuation marks at all. This raises the question whether punctuation should be used for POS tagging at all for these texts.

In order to test the impact of punctuation on tagging performance, three scenarios are considered: tagging with original, modern, and no punctuation marks. In order to provide a fairer comparison, instead of using the supplied parameter file for German, we retrain RFTagger on a prepared set of data. For this purpose, the TIGER corpus (Brants et al., 2002) and version 6 of Tüba-D/Z (Telljohann et al., 2004) are used. First, the two corpora are combined—with minor modifications to the POS tags to make them uniform—and lowercased. The combined corpus has a size of more than 1.6 million tokens. Additionally, for the evaluation without punctuation, a separate tagger model is trained on a version of the TIGER/Tüba corpus where all punctuation marks and sentence boundaries have been removed.

Using these tagger models, tagging performance is evaluated on the gold-standard normalizations with different levels of punctuation. The results are shown in Table 3. For better comparability, accuracy was evaluated excluding punctuation marks in all scenarios.

Tagging with modern punctuation or no punctuation is shown to be best in all cases, with the difference between these two scenarios never being statistically significant ($p > 0.05$). For the Anselm texts, using the original punctuation is worse than using none at all. This is not true for GerManC-GS, though the differences are minor; also, original and modern punctuation are identical for the JubelFeste and Gottesdienst texts, showing that they already follow modern German conventions in this regard.

The results show that removing all punctuation marks does not lead to significant losses in POS tagging accuracy. Indeed, for texts with infrequent and/or inconsistent use of punctuation marks, discarding punctuation is shown to be preferable. For these reasons, the tagging approach without punctuation is used for all following experiments.

4.2 Tagging "with handicaps"

So far, the preprocessing of the historical data includes removing all capitalization and punctuation. Consequently, information about sentence boundaries should also be removed, as it cannot easily be derived from texts without (consistent) punctuation. However, POS tagging with these "handicaps" potentially increases the difficulty of the task in general.

To gauge the extent of this effect, an evaluation on modern data was performed using 10-fold cross-validation on the combined TIGER/Tüba corpus, both with and without these artificial modifications. Table 4 shows the results of

Text	Tokens	Original	Automatically normalized				Gold
			100	250	500	1,000	
Berlin	4,719	28.65%	58.68%	74.89%	75.95%	78.03%	87.07%
Melk	4,550	44.70%	69.63%	74.02%	76.24%	78.66%	87.74%
LeichSermon	2,215	67.95%	72.87%	74.63%	75.85%	78.01%	81.04%
JubelFeste	2,137	82.26%	82.64%	83.62%	86.52%	87.74%	90.03%
Gottesdienst	1,953	88.07%	88.84%	90.27%	91.30%	91.65%	92.27%

Table 5: POS tagging accuracy on texts without punctuation and capitalization, for tagging on the original data, the gold-standard normalization, and automatic normalizations using the first *n* tokens as training data

this experiment; tagging accuracy drops from 96.85% to 95.74% when removing capitalization and punctuation. While this change is significant ($p < 0.01$) considering the corpus size, with regard to the effort involved in manually annotating whole texts with modern capitalization and punctuation marks, it seems small enough to make tagging without this information a viable approach for historical data.

4.3 Tagging historical data

POS tagging on the historical texts is evaluated in three different scenarios: first, tagging on the simplified, but otherwise unmodified, original texts; second, tagging on the gold-standard normalizations; and third, tagging on texts which have been normalized automatically as described in Sec. 3.

For automatic normalization, the first *n* tokens of a text were used for training the Norma tool, with different values for *n* (cf. Sec. 3.2). Only the remainder of the text has then been automatically processed by Norma. This means that, e.g., for a text with 500 tokens used for training, POS tagging is performed on a version of the text consisting of 500 gold-standard normalizations plus automatically generated normalizations for the remainder of the text. This evaluation method models a typical application scenario, where a tradeoff is made between no manual effort (= tagging on the original) and full manual preprocessing (= tagging on the gold-standard).

Full evaluation results are shown in Table 5. Tagging accuracy roughly correlates with normalization accuracy (cf. Table 2); it tends to be slightly above the normalization score for Anselm and a few points below that score for GerManC-GS. Tagging on the original, historical data is particularly inaccurate for the Anselm texts, with the Berlin text only achieving an accuracy of 28.7%. This again highlights the need for specialized tagging methods on such types of data. The GerManC-GS texts from the 18th century perform much better without normalization, with accuracies up to 88% for the Gottesdienst text. These results mainly confirm the observations that the Anselm texts show much more variety in spelling than the newer texts from GerManC-GS.

Similar to the results for normalization, using only 100 tokens for training is enough to increase tagging accuracy for the Melk text from 45% to 70%. For Berlin, this method results in an even higher relative increase, more than doubling the number of correct POS tags. Results for these texts can be improved further to about 74% when using 250 tokens for training; after this figure, POS tagging seems to profit less from increasing the size of the training set, with accuracies around 78% for a training set of 1,000 tokens.

The GerManC-GS texts, particularly JubelFeste and Gottesdienst, do not benefit as much from a small number of training tokens. With 100 tokens, POS tagging accuracy only increases by 0.38–0.77 percentage points. However, these texts already have a comparatively high baseline to start with (82–88%). As they are already much closer to modern German spelling, fewer wordforms have spelling variations at all; consequently, more training data is required to capture a similar amount of variant wordforms as in the Anselm texts. Indeed, when increasing the training portion to 1,000 tokens, the benefit of spelling normalization becomes more pronounced.

Curiously, for the LeichSermon text, even the gold-standard normalization only achieves 81% accuracy, which is significantly lower than for any other text in the evaluation. This is un-

expected, considering that the text is much more recent than Berlin and Melk. The reason for this discrepancy is the frequent use of bible verse numbers in LeichSermon, which are written as numerals followed by a dot and annotated as CARD (cardinal number) in the gold-standard data. In the TIGER corpus and Tüba-D/Z, such numerals are treated as ordinal numbers and tagged as ADJA, leading to a high number of mismatching tags.

4.4 Error analysis

POS tagging results for the historical texts are still considerably worse than those for modern data, even when tagging on gold-standard normalizations (81–92% vs. 95.74%). There are several factors responsible for this.

It is important to observe that even perfectly normalized historical data has different characteristics than modern data, as normalization only affects the spelling of wordforms. One potential source of errors are semantic changes, as shown in Ex. (4) from the LeichSermon text: the wordform *so* is an adverb in modern German, but is frequently used as a relative pronoun (PRELS[2]) in ENHG, which nevers occurs in the training data of the TIGER/Tüba corpus.

(4)

die	*faelle*	*so*	*aus*	*schwacheit*
die	fälle	so	aus	schwachheit
ART	NN	PRELS	APPR	NN
geschehen				
geschehen				
VVPP				

"the cases which occur out of weakness"

Extinct wordforms are a major problem for the normalization approach. They cannot usually be normalized to a modern wordform by applying spelling changes, but would have to be mapped on a word-by-word basis. However, both GerManC-GS and the normalization layer of Anselm[3] map extinct wordforms to artificial lemmas, which are still useful to identify spelling variants, but impractical for this POS tagging approach. A common example in Melk is *czuhant* "immediately", which is mapped to the artificial lemma *zehant*, but would rather be expressed as *sofort* in modern German:

(5)

czuhant	*chust*	*iudas*	*mein*	*chint*
zehant	küsst	judas	mein	kind
ADV	VVFIN	NE	PPOSAT	NN

"Immediately, Judas kisses my child"

Finally, a significant number of errors appears to result from limitations of the modern TIGER/Tüba corpus used to train the POS tagger. This corpus is created from newspaper texts, which are typically written in a rather formal style. The Anselm texts, on the other hand, consist of question/answer sets which contain a lot of direct speech. Similarly, the Gottesdienst text is a religious speech which addresses its audience right from the beginning. Ex. (6) shows a phrase that occurs frequently in the Berlin text:

(6)

sieh	*anselm*
VVIMP	NE

"Look, Anselm"

The imperative form *sieh* "look" is used 24 times in the Berlin text, but typically mistagged as a proper noun (NE) despite being correctly normalized. A look at the TIGER/Tüba training data reveals the cause for this: the wordform *sieh* does not occur there at all; only the standard form *siehe* was learned. Imperative verb forms in general are very uncommon in TIGER/Tüba, only making up 397 tokens (0.02%). In comparison, the gold-standard POS annotation of Berlin already contains 43 imperative verb forms (0.91%).

Similarly, the religious texts in Anselm and GerManC-GS often use vocabulary that is rarely used in newspaper text. Ex. (7) shows the finite verb form *verschmähten* "despised/spurned", which has only one occurrence in the TIGER/Tüba corpus where it was used as an adjective instead, inevitably leading to a tagging error.

(7)

vnd	*vorsmeten*	*yn*
und	verschmähten	ihn
KON	VVFIN	PPER

"and [they] despised him"

These examples show that even if spelling normalization was done perfectly on historical texts, semantic/syntactic variation and domain adaptation of the POS tagger provide further obstacles for achieving higher tagging accuracies.

[2]Actually, GerManC-GS annotates *so* in this example with the new tag PTKREL, which is mapped back to PRELS for reasons of compatibility. As PTKREL is not found in TIGER or Tüba-D/Z, keeping this tag would not solve the problem here, though.

[3]The Anselm corpus provides an additional "modernization" layer which maps extinct forms to actual modern words, but a first evaluation showed that using this layer has a negative impact on overall normalization accuracy.

5 Related work

For automatic spelling normalization, VARD 2 (Baron and Rayson, 2008) is another tool that has been developed for Early Modern English. It has been successfully adapted to other languages, e.g. Portuguese (Hendrickx and Marquilhas, 2011), though previous experiments found it to perform worse than Norma on the Anselm data (Bollmann, 2012). Jurish (2010) presents a normalization method that includes token context, which seems to be the logical next step to further improve normalization results.

POS tagging on normalized data has been tried for the GerManC-GS corpus before with an average accuracy of 79.7% (Scheible et al., 2011a), however, only manual normalization was considered. For English, Rayson et al. (2007) report an accuracy of 89–91% on gold standard normalizations and 85–89% on automatically normalized texts. Hendrickx and Marquilhas (2011) perform a similar evaluation for Portuguese, achieving 86.6% and 83.4% on gold standard and automatic normalizations, respectively.

There are some notable differences, however, between the aforementioned studies and the approach outlined here. Firstly, those studies using automatic normalization methods typically utilize either a much higher amount of training data or some kind of manually crafted resource. VARD, for instance, uses a manually compiled list of spelling variants totalling more than 45,000 entries (Rayson et al., 2005), while Hendrickx and Marquilhas (2011) use a training set of more than 37,000 tokens. While I certainly expect to improve the results in the future by using full texts from the Anselm and/or GerManC-GS corpora as basis for training, this approach might not always be feasible. The approach presented here, requiring only a few hundred tokens for training, seems especially suited for languages where projects to create historical corpora have only been started, and therefore do not have large amounts of previously annotated training material to fall back to.

Secondly, the Anselm texts evaluated here show a much lower baseline than the texts evaluated in other studies. Without normalization, POS tagging accuracy is 82–88% in Rayson et al. (2007), 76.9% in Hendrickx and Marquilhas (2011), and 69.6% for the German data in Scheible et al. (2011a). The texts from Berlin and Melk, on the other hand, perform much worse without the normalization step (28.7% and 44.7%, respectively). This suggests a higher amount of variance in the Anselm data compared to the types of text used in previous studies, making their automatic processing a potentially more challenging problem. Also, annotated data from these studies is less likely to be useful as training data for these texts.

6 Conclusion

I presented an approach to part-of-speech tagging for historical texts that uses spelling normalization as a preprocessing step. Evaluation on texts from Early New High German showed that by manually normalizing 250 tokens of a text and using them as training data, automatic normalization of the remaining text performs well enough to result in a notable increase in POS tagging accuracy. Texts with more spelling variation were shown to benefit more from this approach than texts which are already closer to the modern target language.

For one German manuscript from the 15th century, this method increased tagging accuracy from 28.65% to 74.89%. While this is still far from the accuracy scores reported for modern language data, and also quite a bit worse than tagging on the gold-standard normalization (87.07% for this text), it offers a way to facilitate the (semi-automatic) POS annotation of historical texts with relatively minor effort. Furthermore, as it does not require a sizeable amount of training data, this approach is potentially interesting for less-resourced language varieties in general, assuming some level of graphematic similarity to a well-resourced target language.

Future work should likely consider inclusion of token context for the normalization as proposed by Jurish (2010). Analysis of the POS tagging errors also highlighted some of the problems that remain. Domain-specific differences can negatively impact tagging performance even on perfectly normalized data. Furthermore, spelling normalization cannot account for semantic and syntactic peculiarities of historical language. For a corpus of Old Spanish, this led Sánchez-Marco et al. (2010) to abandon the normalization approach and use a customized POS tagger instead. On the other hand, a study by Dipper (2010) showed that normalization is still beneficial even when retraining a tagger on a corpus of historical data. Future research could try to combine a normalization step with a modified POS tagger to improve the results further.

References

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

Marcel Bollmann, Stefanie Dipper, Julia Krasselt, and Florian Petran. 2012. Manual and semi-automatic normalization of historical spelling – Case studies from Early New High German. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 342–350, Vienna, Austria.

Marcel Bollmann. 2012. (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, Portugal.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In Erhard Hinrichs and Kiril Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, Sozopol, Bulgaria.

Thorsten Brants. 2000. TnT — a statistical part-of-speech tagger. In *Proceedings of ANLP 2000*, pages 224–231, Seattle, USA.

Stefanie Dipper and Simone Schultz-Balluff. 2013. The Anselm corpus: Methods and perspectives of a parallel aligned corpus. In *Proceedings of the NODALIDA Workshop on Computational Historical Linguistics*.

Stefanie Dipper. 2010. POS-tagging of historical language data: First experiments. In *Proceedings of KONVENS 2010*, pages 117–121, Saarbrücken, Germany.

Iris Hendrickx and Rita Marquilhas. 2011. From old texts to modern spellings: an experiment in automatic normalisation. *JLCL*, 26(2):65–76.

Bryan Jurish. 2010. More than words: using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.

Paul Rayson, Dawn Archer, and Nicholas Smith. 2005. VARD versus Word. A comparison of the UCREL variant detector and modern spell checkers on english historical corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK.

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.

Cristina Sánchez-Marco, Gemma Boleda, Josep Maria Fontana, and Judith Domingo. 2010. Annotation and representation of a diachronic corpus of Spanish. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 2713–2718.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011a. 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)*, pages 19–23, Portland, Oregon, USA.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011b. 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, USA.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING '08*, Manchester, Great Britain.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2229–2235, Lisbon, Portugal.