POS-tagging of Historical Dutch

Dieuwke Hupkes, Rens Bod
Institute for Logic, Language and Computation
University of Amsterdam
Science Park 107
D.hupkes@uva.nl, rens.bod@uva.nl

Abstract

We present a study of the adequacy of current methods that are used for POS-tagging historical Dutch texts, as well as an exploration of the influence of employing different techniques to improve upon the current practice. The main focus of this paper is on (unsupervised) methods that are easily adaptable for different domains without requiring extensive manual input. It was found that modernising the spelling of corpora prior to tagging them with a tagger trained on contemporary Dutch results in a large increase in accuracy, but that spelling normalisation alone is not sufficient to obtain state-of-the-art results. The best results were achieved by training a POS-tagger on a corpus automatically annotated by projecting (automatically assigned) POS-tags via word alignments from a contemporary corpus. This result is promising, as it was reached without including any domain knowledge or context dependencies. We argue that the insights of this study combined with semi-supervised learning techniques for domain adaptation can be used to develop a general-purpose diachronic tagger for Dutch.

Keywords: POS-tagging, historical Dutch, parallel corpora

1. Introduction

To extract information from a (historical) text, it is often helpful to know the grammatical categories (or part-of-speech tags) of the words in this text. High-performance automatic part-of-speech taggers (POS-taggers) can be trained when large amounts of annotated training data are available, but automatically POS-tagging low-resource languages for which such data do not exist has proved to be a challenging task. When aiming to automatically tag historical data, one is confronted with an additional difficulty: standardisation of orthography is relatively recent and thus historical corpora often contain a large variation in spelling, which effectively increases the amount of annotated training data necessary to learn a good model.

One approach to address this orthographical variation is to use a respelling tool to normalise/modernise the spelling of a text, prior to tagging it with a tagger trained on modern data of the same language. Rayson et al. (2007) found that for Middle English normalising the spelling of a text increases the accuracy of a rule based modern English tagger from just under 0.82 to 0.85. For manual modernisation, they report an accuracy of 0.89, indicating that to obtain state-of-the-art tagging results for Middle English also lexical and/or syntactical variation should be considered.

Another approach to POS-tagging historical text is to transfer annotation via parallel corpora. Positive results for this technique have been reported for obtaining annotations for closely related languages (e.g., Bentivogli et al. (2004; Van Huyssteen and Pilon (2009; Yarowsky et al. (2001)). Moon and Baldridge (2007) report good results for this method for tagging historical English. However, the applicability of this approach is quite limited, as it requires the availability of a parallel corpus with a similar language for which a good POS-tagger is available.

In this study we focus on POS-tagging 17th-century Dutch texts. As there is little POS-annotated data available for this period, supervised POS-taggers for do not exist.[1] Currently, researchers working with material from this period often resort to POS-taggers trained on contemporary Dutch,[2] although their adequacy for historical texts is highly questionable. A study that evaluates the quality of current annotations and explores methods for improvement is currently lacking.

The goal of this paper is firstly to present a thorough analysis of the adequacy of currently used taggers for historical Dutch and secondly to explore methods for generating higher accuracy tags. In particular, we will asses the effect of different methods for preprocessing (spelling normalisation, as well as word-for-word translation of the text) on the accuracy of tags generated with a tagger trained on contemporary Dutch and we will explore whether making adaptations in the tagger based on knowledge extracted from a diachronic parallel corpus can improve tagging results. We focus on techniques that are simple and easily extendable for different domains. For all methods, we will test the within domain accuracy, but also evaluate the generalisability. Finally, we will discuss how these results can be used in further research to develop methods to automaticall generate taggers for different periods of historical Dutch.

2. Data

Our experiments focus on tagging 17th-century Dutch data. As even within one period there is still a considerable amount of variation, we use 2 texts from 2 different domains: *Iovrnael ofte gedenckwaerdige beschrijvinghe*, a scheepsjournaal (ship's logbook) published in 1646 (Bontekoe, 2013) and the Dutch Bible translation of 1637 (Statenbijbel, 2008).

[1] Supervised taggers/lemmatizers for medieval Dutch exist (Kestemont et al., 2014; van Halteren and Rem, 2013).

[2] See, e.g., `http://www.nederlab.nl/`.

ADJ	adjective	BW	adverb
LID	article	N	noun
TSW	interjection	TW	numeral
VG	conjunction	VNW	pronoun
VZ	preposition	WW	verb
SPEC(e)	proper name	SPEC(v)	borrowed
LET	punctuation		

Table 1: Coarse POS-tags used for annotating test corpora. Used tagging conventions can be found in Van Eynde (2004)

2.1. Test corpora

For testing, we manually annotate 50 random sentences from both corpora with coarse POS-tags (13 in total, see Table 1). We use the tagset from Corpus Gesproken Nederlands (Oostdijk, 2002), as well as their tagging conventions (Van Eynde, 2004). We use the Bible corpus (**Bible1637**, 1368 tokens) for development and testing, and the Bontekoe corpus (**Bontekoe**, 1565 tokens) to test the generalisability of our results to other domains. For comparison, we also annotate the more modern translation of the 50 **Bible1637** sentences that can be found in the Dutch Bible translation of 1977.

2.2. Diachronic Parallel Corpus

The rest of the two Bible texts (31172 lines, over 900000 tokens per text) we use as a diachronic parallel corpus. We lowercased the two texts and employed a machine translation tool[3] (5 iterations for both models) to align the sentences on the word level, resulting in largely monotone alignments. Fig. 2 shows an example of such an alignment. A quick inspection shows that the resulting word alignments contain mistakes, but are generally of high quality.

2.3. Letters as Loot

A third dataset we have available is the Letters as Loot corpus (van der Wal et al., 2012), a dataset consisting of 1000 letters (over 40.000 tokens) written by sailors between the second half of the 17th century and the beginning of the 19th century. The POS-annotation of the corpus is checked manually and thus of high quality, but both the conventions for tagging and the set of labels differ from the CGN tagset; this renders the corpus suboptimal for training and testing purposes (for the present study). Nevertheless, we will use the corpus to increase the vocabulary of a tagger in a later stadium.

3. Taggers

We use two different taggers: a memory based tagger called MBT (Daelemans et al., 2010), trained on a contemporary Dutch corpus with over 11 million annotated words and Trigram'n'Tags (Brants, 2000), a very efficient hidden-markov model tagger that does not come with a pretrained model for Dutch but can be trained easily on an annotated dataset.

[3]The Berkeley Aligner `https://code.google.com/p/berkeleyaligner/`

To disambiguate the tags of words seen in the training corpus, MBT uses context information from both the left and right side of the word. To assign tags to words unknown to the tagger, additional features are used such as the first and last letters of the focus word and whether the word contains capital letters or numbers.
Trigrams'n'Tags (TnT) is a trigram-based tagger, whose parameters are estimated from a corpus and then smoothed using a context-independent variant of linear interpolation. The interpolation parameters are estimated by deleted interpolation. Unknown words are tagged based on suffix analysis and a flag indicating whether the focus word is capitalised.

4. Experiments

To obtain a baseline, we tag all test sets with MBT and evaluate the average tagging accuracy per word, ignoring punctuation tags. For the contemporary corpus **Bible1977** we find an accuracy of 0.96, which is slightly lower than the accuracy reported in Van Eynde (2004). This discrepancy may be caused by ignoring the punctuation tags (which are always correct), but is most likely also partly caused by the slightly archaic language use in the corpus. The tagging accuracy of the historical datasets is low, around 0.60 (see Table 2)

4.1. Respelling

An analysis of the confusion matrix of the tags assigned to the historical corpus shows that a large part of the mistakes is due to divergences in spelling. For instance, many words are assigned the tag 'SPEC(v)', which is used for words that are considered to be not morphosyntactically integrated in the language. As the tagger uses statistics of low frequent words in the training corpus to tag unknown words in the test corpus, the unknown-word module systematically fails to classify words in high frequent closed categories (such as pronouns and conjunctions) whose spelling diverges from the spelling in the training corpus. Furthermore, the irregular capitalisation impedes feature based classification.
Applying a small set of simple rewrite rules that accounts for systematic cases (such as the change from "ae" to "aa") and takes care of respelling most closed-class words (such as pronouns) leads to a significant improvement of around 15 percentage points (see Table 2). We can interpret this result as a lowerbound for the improvement that is easily achievable through simple adaptations in spelling.
There is a vast amount of research on automatic spelling normalisation (e.g., Hendrickx and Marquilhas (2011); Reynaert (2011); Reynaert et al. (2012)) and modernisation (e.g., Rayson et al. (2005); Koolen et al. (2006)). To assess the potential usability of such respelling tools for this intent, we also determine an estimate of the *upperbound* of the results that can be achieved by spelling-based approaches by manually modernising the spelling of all words in the test corpora. To get a more realistic upperbound, we aim to modernise spelling but preserve lexical and syntactical differences (such as the change of the meaning of the word "en" from "not" to "and""). We find an accuracy of 0.89 for the **bible1637** corpus and 0.82 for the Bontekoe corpus (see Table 2).

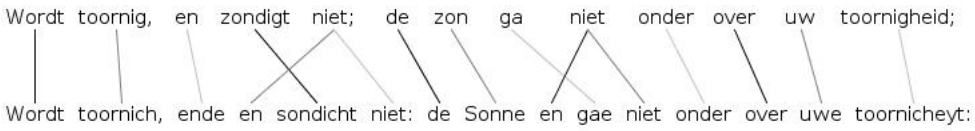

Figure 1: Example word alignment

Figure 2: Example word alignment from our corpus (the visualisation tool used to generate this picture is available at `https://bitbucket.org/teamwildtreechase/hatparsing/`). Note that although in this example every word aligned is with exactly one other word, this is not necessarily always the case.

	MBT tags	Rewrite Rules	Manual Respelling
Bible1637	0.61	0.74	0.89
Bontekoe	0.60	0.73	0.82

Table 2: Influence of respelling prior to tagging with contemporary Dutch tagger. Average accuracy per word.

4.2. Using Information from a Parallel Corpus

Since 17th-century Dutch is similar to contemporary Dutch, both on a lexical and syntactical level (spelling differences aside), it seems plausible that the resources for modern Dutch can be used to bootstrap a tagger for historical Dutch. We first conduct an experiment to determine whether we can employ the alignment information from a parallel corpus to learn how to modernise/normalise the spelling in historical corpora, and consequently investigate whether a tagger can be generated by a training corpus annotated by transferring information via word alignments.[4] Note that our aim is not to evaluate how well annotation can be tranferred via word alignments (see for instance Van Huyssteen and Pilon (2009); Bentivogli et al. (2004); Hwa et al. (2002)), but rather to employ this information to tag other texts, such that the results are not restricted to historical texts for which parallel corpora are available.

4.2.1. Learning to translate words

In our first experiment, we infer a dictionary from our word-aligned parallel corpus by matching every word with the word it was most often aligned with.[5] The resulting dictionary contains 24078 entries (some of which are names). We use the dictionary to replace in the test set either every word for which a 'translation' is available, or only words that did not occur in a list of 'modern' words that occurred in the rest of the 1977 Bible. We use the same procedure for the out of domain Bontekoe corpus. In the **bible1637** corpus, 529 out of 1370 tokens were replaced in the in the latter condition and 713 in the former; 13 tokens in the test set were not available in the dictionary or in the known word list. The Bontekoe corpus contained many more unknown words: 322. Out of 1564 tokens, 478 and 339 were replaced in the former and latter condition, respectively. Note that the former condition - in which all possible words are replaced - can be interpreted as a rough word-for-word translation of the text.

[4] A similar experiment was conducted by Moon and Baldridge (2007).

[5] This is the simplest way in which this could be done, it does not take into account any context dependencies.

We tag both versions of the test corpora with MBT and evaluate the results. The results on the within-domain bible corpus are good (an accuracy of 0.92, which is higher than the strictly spelling based upperbound we determined previously), but do not generalise well to the Bontekoe corpus (an accuracy of around 0.80, see Table 3).

	Baseline	Replace Unknown	Replace All
Bontekoe	0.60	0.80	0.78
Bible1637	0.60	0.90	0.92

Table 3: Influence of replacing words using a dictionary inferred from a parallel corpus. Average tagging accuracy per word.

4.2.2. Training a new tagger

A different approach to improve tagging results for historical texts is to adjust the paramaters of the tagger, rather than preprocessing the text prior to tagging it. The most obvious way of doing this is to retrain a tagger on a corpus with data more similar to the training data. However, to retrain a tagger, a fairly large (annotated) training corpus is required. We investigate if such a training corpus can be created from a diachronic parallel corpus by projecting (automatically generated) tags via a word alignment from the contemporary to the historical side of the corpus. To project the tags, we follow a simple protocol:

1. Every token in the 1637 corpus that is aligned with only one token in the 1977 corpus (264714 tokens) will be assigned the tag of that token;
2. Every token in the 1637 corpus that is aligned with two tokens with tags X and Y (2751 tokens) will be assigned the tag X+Y if X $\neq$ Y, or X otherwise;
3. The tokens that then are not assigned a tag after step 1 and 2 will be assigned the tag that they are most often associated with in the corpus (8000 tokens);
4. The rest of the tokens ($<$150) are manually tagged using regular expressions.

To estimate the accuracy of the tagged corpus **bible1637annotated** we apply the same procedure to the entire corpus (including the test set) and evaluate the accuracy of the tags of the test set, which is around 0.92.
We train TnT on the automatically annotated training corpus (we use the default settings for training) and use the resulting tagger to tag our test corpora. In a post-processing

	Training corpus		
Corpus	Bible1637annotated	Letters as Loot	Bible1637annotated + Letters as Loot
Bible1637	0.94	0.85	0.92
Bontekoe	0.74	0.81	0.84

Table 4: Retraining a tagger on an annotated historical training corpus. Average tagging accuracy per word.

step we replace the tag 'LID+VZ', that did not occur in the gold standard, with the tag 'LID'. The accuracy of the resulting tagger on the **bible1637** corpus is high (0.94) but the results do not transfer well to another domain: the accuracy on the Bontekoe corpus is only 0.74.

Studying the confusion matrix of the tags of the Bontekoe corpus, we find that no class of words is systematically tagged well. The fact that even high frequent words such as articles and numerals are regularly assigned the wrong tag leads to the impression that adding more training data could be beneficial. To test this hypothesis, we train a tagger on a combined corpus consisting of the **bible1637annotated** corpus and the Letters as Loot corpus. For comparison, we also trained a tagger on the Letters as Loot corpus without the **bible1637annotated** data.

The results in Table 4 show that adding more material does indeed have a positive effect on the outside domain results of the tagger, albeit while having a small negative effect on the within domain results.

5. Discussion and Future Work

Our experiments show that POS-taggers for contemporary Dutch texts are not suitable for historical data, generating tags with an accuracy of around 0.60 (see Table 2). We aimed to investigate techniques that were very simple and generally applicable, and do not require extensive amounts of manual work to tailor to different domains.

We confirmed the findings of Rayson et al. (2007) that there appears to be a ceiling to the improvement that can be achieved by spelling based approaches. Even with manual modernisation of spelling, the tagger does not achieve an accuracy higher than 0.90, which indicates that achieving state-of-the-art results on 17th-century Dutch texts requires more than a clever respelling algorithm. However, further improvement for preprocessing based approaches seems possible if also lexical variation is taken into account. Using a dictionary derived from a parallel corpus - which finds a *translation* for the words in the corpus, rather than merely respelling them - results in an accuracy of 0.91 for within domain text, but does not generalise well to different domains (see Table 3). This result suggest that more sophisticated word-for-word translation methods, or the use of manually created dictionaries that map historical wordforms to modern lemma's[6] could lead to further improvements.

Another direction that can be taken is to develop a tagger that is better tuned to historical material. We tested if a tagger could be trained on a corpus automatically annotated with POS-tags projected via word alignments. This gives excellent result for a corpus of the same domain as the training corpus (an accuracy of 0.94), but not on a corpus of a different domain (see Table 4). An analysis of the mistakes shows that the drop in accuracy is mostly caused by the fact that many of the words in the out of domain corpus are unknown to the tagger. Adding data from the Letters as Loot corpus to enhance the lexicon (partly) solves this problem, increasing the out of domain accuracy to 0.84.

We surmise that to improve upon these results, the focus should be on developing methods for domain adaptation (similar to for instance (Yang and Eisenstein, 2015; Yang and Eisenstein, 2016)) For instance, using other sources (e.g., (Instituut voor Nederlandse Lexicologie, 2007; INL, 2015)) to add more words to the lexicon of the tagger is likely to be beneficial. Another approach could be to use semi-supervised approaches to fine-tune the parameters of the retrained tagger after adding entries for unknown words of the testset (see for instance (Deoskar et al., 2013)). Baum-Welch re-estimation of parameters has shown to be very strongly dependent on initialisation (Elworthy, 1994; Merialdo, 1994), but has the potential of finding reasonable solutions given a good start (Goldberg et al., 2008). Using information from parallel corpora and previously mentioned manual sources could be used to find such a starting point. To decrease sparsity of parameters, this approach could be combined with a preprocessing step in which spelling is normalised. An advantage of this approach is that it could provide an automised way to learn taggers for different domains of historical texts. If successful, similar techniques can also be used to tackle lemmatisation of historical texts, as well as tagging of other historic languages.

Although orthographic variation might hinder their applicability, a third line of reseach that could be considered for tagging (low resource) historical texts is research on semi- or unsupervised POS-tagging, such as (Deoskar et al., 2013; Garrette and Baldridge, 2013; Goldberg et al., 2008; Brill, 1995; Goldwater and Griffiths, 2007). (Yang and Eisenstein, 2015; Yang and Eisenstein, 2016) have reported good results for unsupervised tagging of historical English.

6. Conclusion

We studied several methods for assigning POS-tags to historical Dutch texts from a period for which little annotated data are available and ortography was not yet standardised. We confirmed that POS-taggers trained on contemporary Dutch are not adequate for tagging 17th-century Dutch corpora, and explored different techniques to improve upon their tagging accuracy of around 0.60. We showed that respelling algorithms are effective, but not sufficient to obtain state-of-the-art POS-tagging results. The largest improvements were obtained by retraining a POS-tagger on an automatically annotated historical corpus. The improvement

[6] E.g., Woordenboek der Nederlandse Taal (Instituut voor Nederlandse Lexicologie, 2007)

	Baseline	Rewrite	Translation		Retrain tagger on		
			all	unknown	Bible1637annotated	Letters	**Bible1637annotated**
Bible1637	0.60	0.74	0.78	0.80	0.94	0.85	**0.92**
Bontekoe	0.60	0.73	0.92	0.90	0.74	0.81	**0.84**

Table 5: Summary of results.

subsists across domains, but the within domain results (an accuracy of 0.94) are significantly better than for other domains (0.84 accuracy). However, the results are a tremendous improvement - of 34 and 23 percentage points, respectively - over the baseline accuracy of currently used taggers. Notably, none of the methods explored were tailored to a specific domain. We chose to not make small adaptations to the tagger based on our knowledge about the corpus, even though that could have led to further improvements. In the future, we will focus on finding these adaptations automatically, by combining the techniques discussed in this paper with semi-supervised learning paradigms. We argue that the results of the current study constitute a step towards developing a general-purpose diachronic tagger for Dutch, and can also be applied to other languages and tasks.

7. Bibliographical References

Bentivogli, L., Forner, P., and Pianta, E. (2004). Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, page 364. Association for Computational Linguistics.

Brants, T. (2000). Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics.

Brill, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. In *Proceedings of the third workshop on very large corpora*, volume 30, pages 1–13. Somerset, New Jersey: Association for Computational Linguistics.

Daelemans, W., Zavrel, J., Van den Bosch, A., and Van der Sloot, K. (2010). Mbt: memory-based tagger. *Version*, 3:10–04.

Deoskar, T., Mylonakis, M., and Sima'an, K. (2013). Learning structural dependencies of words in the zipfian tail. *Journal of Logic and Computation*.

Elworthy, D. (1994). Does baum-welch re-estimation help taggers? In *Proceedings of the fourth conference on Applied natural language processing*, pages 53–58. Association for Computational Linguistics.

Garrette, D. and Baldridge, J. (2013). Learning a part-of-speech tagger from two hours of annotation. In *HLT-NAACL*, pages 138–147. Citeseer.

Goldberg, Y., Adler, M., and Elhadad, M. (2008). Em can find pretty good hmm pos-taggers (when given a good start). In *ACL*, pages 746–754. Citeseer.

Goldwater, S. and Griffiths, T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *Annual meeting-association for computational linguistics*, volume 45, page 744.

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

Hwa, R., Resnik, P., Weinberg, A., and Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 392–399. Association for Computational Linguistics.

Kestemont, M., de Pauw, G., van Nie R., and Daelemans, W. (2014). Towards a general purpose tagger-lemmatizer for pre-modern Dutch. Conference talk presented at the Digital Humanities 2014 Benelux Conference.

Koolen, M., Adriaans, F., Kamps, J., and De Rijke, M. (2006). A cross-language approach to historic document retrieval. In *Advances in Information Retrieval*, pages 407–419. Springer.

Merialdo, B. (1994). Tagging english text with a probabilistic model. *Computational linguistics*, 20(2):155–171.

Moon, T. and Baldridge, J. (2007). Part-of-speech tagging for middle english through alignment and projection of parallel diachronic texts. In *EMNLP-CoNLL*, pages 390–399.

Oostdijk, N. (2002). Het Corpus Gesproken Nederlands.

Rayson, P., Archer, D., and Smith, N. (2005). Vard versus word: A comparison of the ucrel variant detector and modern spellcheckers on english historical corpora.

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.

Reynaert, M., Hendrickx, I., and Marquilhas, R. (2012). Historical spelling normalization. a comparison of two statistical methods: Ticcl and vard2. *on Annotation of Corpora for Research in the Humanities (ACRH-2)*, page 87.

Reynaert, M. W. (2011). Character confusion versus focus word-based correction of spelling and ocr variants in corpora. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(2):173–187.

van der Wal, M., Rutten, G., Simons, T., et al. (2012). Letters as loot. confiscated letters filling major gaps in the history of dutch.

Van Eynde, F. (2004). Part of speech tagging en lemmatisering van het corpus gesproken nederlands. *KU Leuven*.

van Halteren, H. and Rem, M. (2013). Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century dutch charters. *Language resources and evaluation*, 47(4):1233–1259.

Van Huyssteen, G. B. and Pilon, S. (2009). Rule-based conversion of closely-related languages: a dutch-to-afrikaans convertor.

Yang, Y. and Eisenstein, J. (2015). Unsupervised multi-

domain adaptation with feature embeddings. *Proc. of NAACL-HIT*.

Yang, Y. and Eisenstein, J. (2016). Part-of-speech tagging for historical english. *arXiv preprint arXiv:1603.03144*.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

8. Language Resource References

W.IJ. Bontekoe. (2013). *Journael ofte gedenckwaerdige beschrijvingen van de Oost-Indische reijse*. dbnl, G.J. Hoogewerff.

INL. (2015). *Computationeel Historisch Lexicon*. http://www.inl.nl/onderzoek-a-onderwijs/lexicologie-a-lexicografie/wnt.

Instituut voor Nederlandse Lexicologie. (2007). *Woordenboek der Nederlandse Taal (WNT)*. http://www.inl.nl/onderzoek-a-onderwijs/lexicologie-a-lexicografie/wnt.

Statenbijbel. (2008). *Biblia, dat is: De gantsche H. Schrifture (statenvertaling van 1637)*. dbnl.