Improving Part-of-Speech Tagging of Historical Text by First Translating to Modern Text

Erik Sang

Improving Part-of-Speech Tagging of Historical Text by First Translating to Modern Text

Erik Tjong Kim Sang

Meertens Institute Amsterdam
erik.tjong.kim.sang@meertens.knaw.nl

Abstract. We explore the task of automatically assigning syntactic tags (known as part-of-speech tags) like *Noun* and *Verb* to words in seventeenth-century Dutch text. Tools exist for performing this task for modern texts but they perform poorly on historical texts because of language changes. We test several methods for translating the words in the historical text to modern equivalents before applying the tag assignment tools. We show that this additional translation step improves the quality of the automatic syntactic analysis. Further improvements are possible when the lexicons and text collections used for developing the translation process, are extended in size.

1 Introduction

Nederlab[1] [3] is a large-scale effort to provide to the research community digital versions of texts of the past millennium that are written in Dutch. In order to enable various types of linguistic and historical research, the project aims at providing linguistic annotations with the text. Because of the large volumes of text that are involved, most of the annotations will be generated by automatic tools. Most of the present tools have been built to process twentieth century newspaper text. Since the language Dutch has changed considerably in the last centuries [21], the tools perform poorly on historical texts.

There are two ways to improve the quality of automatic linguistic annotation. One is to retrain the tools on historical texts [16]. However, this requires that gold standard training data are created for all relevant linguistic annotation tasks *and* for all relevant time periods: an enormous task. The second method relies on translating the historical texts to a modern variant which can then be processed by the available language processing tools. The expected quality of the annotations will be somewhat lower than of the special purpose tools but the gold standard annotation requirements for this approach are more manageable.

In this paper, we explore the preprocessing method of text translation with the goal of improving the quality of linguistic annotation tools on processing historical text. We focus on one language (Dutch), one time period (the seventeenth century) and one annotation task (assigning syntactic part-of-speech

[1] nederlab.nl

tag to words). We examine four different methods for translating the texts and present a comparison of their effects on tagging quality.

A good quality of the syntactic analysis of texts is important for being able to find specific words in texts. For example, if a linguist or a historian wants to know which historical texts use the verb *ship*, correct syntactic annotation will speed up the search process tremendously. When texts have no syntactic annotation, the researcher must examine many irrelevant documents before finding the ones that he or she needs. The same is true when the syntactic annotations contain errors, when noun versions of the search word are often incorrectly annotated as verb. And when verb occurrences of *ship* have mistakenly been annotated as noun, these will not be found by a search process looking for *ship* tagged as verb.

The text translations produced in our work, can be used for other purposes than improving syntactic annotation. However, for us they serve only this purpose. We will not use the translations for replacing the original texts. The translations may also not be what people expect of them. Translation is a difficult process to automate. For example, the Dutch historical word *beroert* has several equivalents in modern language, like *touched*, *hit* and *sick*. For our purposes, it is not important that the translation process finds the word with correct meaning, but only that the syntactic class of the translation is the same as the one of the historical word (in this case: adjective).

After this introduction, we discuss related work in section two. In section three we describe the four methods we have used for translating historical text to modern Dutch. Section four presents the evaluation results of these approaches for part-of-speech tagging. In section five we conclude.

2 Related work

The field of domain adaptation [15, 9] deals with the problem of applying and improving language tools for text types they have not been developed for. Such adaptations are not only necessary for texts written in older variants of languages but also for texts from different genres, as McClosky et al. [13] shows, with a 30% performance drop for a parser processing text from an out-of-training-domain topic. Recently there have also been attempts to create tools that perform well in different text domains [19].

Archer et al. [1] developed VARD (VARiant Detector), a tool which can be used to convert historical texts to a modern version with standardized spelling, which can then be processed by language tools trained on modern text. The tool has been applied to Early Modern English [20] and to seventeenth-century Dutch [26]. Hupkes and Bod [6] explored semi-supervised learning for tagging historical Dutch texts. Reynaert [17] developed TiCCl, a tool for normalizing Dutch texts by performing automatic spelling correction. The program Adelheid has specifically been developed for lemmatizing and tagging fourteenth-century Dutch [16].

Het eerste Capittel.
translation: *Het eerste hoofdstuk.*
'De Mensch, het edel dier, by Godes hant geschapen.
translation: *'De Mens, het edel dier, door Gods hand geschapen.*
Was, om in stage jeught, sijn lust te mogen rapen;
translation: *Was, om in stage jeugd, zijn lust te mogen rapen;*
Was in het schoon prieel: en waer hy immer ging,
translation: *Was in het mooie prieel: en waar hij steeds ging.*
Daer was hy aengesien als heer van alle ding:
translation: *Daar werd hij gezien als heer van alle dingen:*
Hy vont een schoon gesicht alwaer de boomen groeyden,
translation: *Hij vond een mooi uitzicht waar de bomen groeiden,*
Hy vont een soet geluyt alwaer de beken vloeyden,
translation: *Hij vond een zoet geluid waar de beken vloeiden,*

Fig. 1. Example of seventeenth-century Dutch: the first seven lines of the book *Schat der Gesontheyt* by Johan van Beverwijck, edition 1663 with their translation to modern Dutch (*italic characters* preceded by **translation:**) [12]. The historical Dutch is similar enough to modern Dutch to be understood by speakers of contemporary Dutch but different enough to cause problems for processing by language tools.

Social media messages suffer from the same variety in spelling as historical texts. Techniques used for converting such messages to standardized spelling [5, 10], can also be applied to text from centuries ago.

3 Translation methods

We examine four different methods for translating seventeenth-century Dutch to modern Dutch. An example of the translation task can be found in Figure 1. The two language variants are quite similar but the differences are large enough to create problems for language processing tools.

3.1 Machine translation

Converting seventeenth-century Dutch to modern Dutch is similar to translating one language to another. General machine translation systems have been developed exactly for this purpose. In order to get an idea of the performance of these systems, we applied a state-of-the-art general purpose machine translation system to this task: Moses [11]. We trained Moses with two versions of the Dutch bible Statenvertaling, one from the year 1637 [22] and one from 1888 [23]. After the training phase, Moses required extra data for the the tuning phase. For this purpose, we used the first chapter of the book *Schat der Gesontheyt* by Johan van Beverwijck from 1663 and its translation to contemporary Dutch [12].

We tested the quality of the translation system by applying it to the first 100 lines of the second chapter of the book by Van Beverwijck. We used BLEU [14]

Historical variant	Lemmas
ende	**en** (57642) einde (318) eend (0)
de	**de** (41141) doen (1658)
van	**van** (22251) vinden (160)
het	**het** (21009) hebben (3018)
den	de (41141) **den** (11521)
in	en (57642) **in** (14785)
hy	**hij** (9498) hei (0)
die	**die** (11676)
dat	die (11676) **dat** (9629)
tot	**tot** (10878) totten (0)

Fig. 2. Modern lemmas of seventeenth-century Dutch words, provided by the Lexicon Service of the Dutch Institute for Lexicography [8]. The number of times each lemma occurs in a modern bible version [23] is mentioned between brackets. The lemmas chosen as best modern variant for the historical words are shown in **bold face**: either the modern lemma which is identical to the historical word (like the pair *de-de*) or the modern lemma with the highest bible frequency if no identical variant is available (like the pair *hy-hij*).

as evaluation metric. The translation made by Moses achieved a BLEU score of 0.283 when compared with the gold standard translation by Koomen [12]. We did not know how to interpret this number, so we performed an additional manual translation of the text and compared it with the gold standard. The manual translation obtained a BLEU score of 0.345 while the original text from 1663 was assigned a BLEU score of 0.124. These two scores can be seen as upper and lower bound for the machine translation performances on the test text. The BLEU score achieved by Moses was closer to the upper bound than to the lower bound. In fact, we later found out that no other automatic method evaluated by us achieved a better BLUE score than Moses on the test texts (see Table 1).

Moses requires a large number of computer resources, taking about one hour to start at our machine (Mac mini, 2.3 Ghz, 4Gb). Furthermore, during translation Moses sometimes inserts or deletes words. While this may be the best option for achieving good translations, it complicates the task of linking annotations of modern words back to the original historical words. Therefore we looked for alternative translation methods, that required fewer computational resources and preserved the word positions.

3.2 Historical lexicons

The Dutch Institute for Lexicography (INL) offers the Integrated Language Bank (GTB) [7], an online collection of historical dictionaries, with links of historical words to their modern counterparts [8]. The lexicon service makes it possible to retrieve modern lemmas for historical words. As a test, we retrieved the modern lemmas for the words that occurred five times or more in the Dutch Statenvertaling bible version of 1637 [22]. This resulted in a list of 8,563 words. About

20% of these words were not found in the lexicon, most of these were proper nouns (names). These words were left untranslated. Some of the words had several alternative lemmas assigned to them, see Figure 2. If this was the case, we chose the modern lemma which was equal to the historical word, if one was present among the alternatives, and otherwise we chose the lemma which was most frequent in the modern (1888) version of the Statenvertaling bible [23]. All together 3,948 words (46%) were mapped to themselves. Translating the test text, the first 100 lines of the second chapter of *Schat der Gesontheyt* (1663) by Johan van Beverwijck [12], with the 8,563-word lexicon resulted in a text with a BLEU score of 0.191, better than the baseline of 0.124 but worse than the score of Moses (0.283, see Table 1). However, it should be taken into account that the fact that this method produces lemmas rather than words as present in the gold standard translation, negatively influences its BLUE score.

3.3 Extracting lexicons from parallel text

The translation lexicon used in the previous section links historical words to modern lemmas. The lemmas can be used to determine coarse part-of-speech tags like noun and verb. However, finer distinctions like plural versus singular cannot be derived from the lemmas because the required morphological clues have been lost in the translations process. We do not have parallel word-to-word translation lexicons available but we do have a large parallel text from which such a lexicon could be derived. For this purpose, we examined K-vec, a method for extracting lexicons from unaligned text, put forward by Fung and Church [4]. It divides the text and its translation in K parts of approximately equal size and constructs binary word vectors of length K which state in which parts of the original text or the translated text the word is present. Translated words are likely to appear in the related contexts and thus have similar vectors. We used the Dutch Statenvertaling bible versions of the years 1637 [22] and 1888 [23] as a training corpus. The texts were sentence-aligned so we could use sentences as parts. There were 37,100 sentences in the corpus and from these K-vec identified 16,201 word translation pairs. With the resulting parallel lexicon, we generated a word-by-word translation of the test text *Schat der Gesondheyt* [12]. The translated text proved to be better than the baseline: BLEU score 0.219 compared to 0.124 (see Table 1).

3.4 Orthographic rules

In the evolution of Dutch over the past centuries, changed spelling of words often could be expressed in orthographic rules. Some examples of this can be found in the text in Figure 1: *groeyden* became *groeiden* and *vloeyden* became *vloeiden*. The orthographic rule y⇒i could have been used here to generate the modern variant from the historical variant but the context could also have played a role. In order to find reliable orthographic rules, we collected all sequences of one, two and three characters from words in the 16,201-word learned lexicon from the previous section as well as their possible corresponding translations. We

Frequency	Precision	Rule	Frequency	Precision	Rule
895	0.903	y ⇒ i	201	0.971	dt$ ⇒ d
623	0.967	ae ⇒ aa	161	0.953	ch$ ⇒ g
346	0.989	uy ⇒ ui	99	0.980	gh$ ⇒ g
222	0.996	aen ⇒ aan	11	0.917	ph$ ⇒ f
221	0.978	hey ⇒ hei			
177	0.947	uyt ⇒ uit			
162	0.982	aer ⇒ aar			
150	0.993	∧uy ⇒ ui			
139	0.993	∧ae ⇒ aa			
107	0.930	ck$ ⇒ jk			

Fig. 3. Examples of length-preserving orthographic rules (left) and length-reducing orthographic rules (right). ∧ is used as a start-of-word character indicating that the character substitution can only be made at the beginning of a word. $ is the end-of-word character used in a similar fashion. The derivation method found no length-increasing rules.

restricted the word pairs to those where the translation had the same number of characters as the historical word (8,724 word pairs), where the translation was one character shorter (3,969) or where the translation was one character longer (682). Two additional characters were added to each word: a start-of-word character at the beginning and an end-of-word character at the end so that specific rules dealing with the start and the end of a word could be expressed [24].

We collected all non-identical pairs of character strings which occurred at least ten times in the learned lexicon and in which the historical part corresponded with the same translated part in at least ninety percent of the cases. This resulted 86 rules of pairs of the same length, four rules which removed a single character and no rules that added a character (see Figure 3). When the 90 rules were applied to the test text *Schat der Gesontheyt*, [12] it resulted in a translation with a BLEU score of 0.160. This is the lowest score of the four evaluated translation methods. However, the rule set was able to improve both the learned lexicon output (from 0.219 to 0.229) and that of the historical lexicon (from 0.191 to 0.198) when combined with these methods as a post-processor.

4 Part-of-speech tagging evaluation

In order to evaluate the performance effect of the four translation methods discussed in the previous section, on part-of-speech (POS) tagging, we needed gold standard seventeenth-century POS tagged text. Hupkes [6] supplied us with the two texts she used for evaluating her seventeenth-century POS tagger: a selection of the shipping logbook *Journael ofte gedenckwaerdige beschrijvinge* from the year 1646 by Willem IJsbrantsz Bontekoe (1565 tokens) and a selection of the Dutch bible *Statenvertaling* from the year 1637 (1370 tokens). The texts

were annotated with POS tags according to the CGN format [25]. Only coarse tags were used which means that the tags expressed the differences between POS classes like verb, adjective and noun, but not the subtle distinctions within these classes, like plural versus singular or past versus present tense.

For tagging the texts, we used Frog [2], a state-of-the-art POS tagger for modern Dutch. The tagger assigns fine-grained POS tags to words. Only the coarse-grained parts of these tags have been used for evaluation. It would have been useful to see which words the tagger did not know: we could then have focused the translation process on these words. Unfortunately, the tagger did not report which words were unknown. Frog did report a tag confidence score for each word but it was not possible to derive from this if the input word was known or not.

Before the evaluation of the four translation methods, we tested three alternative part-of-speech tagging approaches. First, we tagged the two seventeenth-century texts with the modern tagger without any text modification. This resulted in accuracy scores of 68.2% for the Bontekoe text and 63.7% for the Statenvertaling text. Clearly there is much room for improvement as Frog is reported to achieve an accuracy of 98.6% on this task for modern text [2].

In order to assess the upper performance ceiling of the translation approaches, we tagged manually created word-by-word translations of the two texts. This time we achieved accuracy scores of 88.8% for the Bontekoe text and 91.2% for the Statenvertaling text. The latter score shows that the translation approach is potentially as strong as retraining a tagger: Hupkes and Bod [6] reported that their retrained POS tagger achieves accuracies of 84% on the Bontekoe text and 92% on the Statenvertaling texts. Although human translation is expensive and infeasible for the corpus sizes we aim to process, it is an option we consider for smaller texts for which high quality POS tags are required.

As a third alternative method, we evaluated the performance of the Adelheid tagger [16], a tagger which was specifically trained for processing Middle Dutch (1200-1500). Adelheid does not use the CGN tag set but it was possible to convert its output tags to the CGN format. Adelheid performed slightly better than the baseline for the Bontekoe text (71.4% compared to 68.2%) and a lot better for the Statenvertaling text (82.9% compared to 63.7%). It is unclear what is causing this performance difference.

Next, we processed the Bontekoe text with three of the four translation methods and sent the results to the Frog parser. The machine translation system Moses was not used for translation because it does not preserve the word order of the original text and we had no automatic method for linking the POS tags back to the original words. The Statenvertaling text was not used for this part of the evaluation. All four methods have used the Statenvertaling in one way or another in their development phase and therefore evaluation scores on this text would be unreasonably high.

The POS accuracies of the three methods can be found in Table 1. The translations produced with the historical lexicon proved to be as useful for POS tagging (82.0%) as the translation produced with the learned lexicon (82.1%).

	Gesontheyt	Bontekoe	Statenvertaling
Method \ Measure	**BLEU**	**POS accuracy**	**POS accuracy**
Modern tagger	0.124	68.2%	63.7%
Historical tagger	0.124	71.4%	82.9%
Manual translation	0.345	88.8%	91.2%
Machine translation	0.283	NA	
Historical lexicon	0.191	82.0%	
Learned lexicon	0.219	82.1%	
Orthographic rules	0.160	73.4%	

Table 1. Evaluation scores for the three alternative approaches and the four translation methods. Part-of-speech (POS) tagging accuracies concern base tags only. No POS scores are available for the machine translation method because of the difficulty of linking POS tags from its output back to the original words. The Statenvertaling was used as training material for the translation methods and therefore these methods have not been evaluated on those data.

This shows that translating to lemmas rather than to words is a valid approach for improving the quality of assigning coarse-grained POS tags. For fine-grained POS tags, lemmas will be insufficient since the morphological clues required for such tags are lost in the translation process. The orthographic rules performed considerably worse than the two lexicon methods (73.4%). Like in the tests in section 3.4, they improved the BLUE scores of the lexicon methods when combined with them as a post-processor. However the associated POS accuracy scores did not improve (82.0% and 81.9%, respectively).

While the two lexicon-based translation methods offer a considerable increase in POS tagging accuracy, there is still some room for improvement in comparison with the manual translation. One way of achieving further improvement would be use a larger historic lexicon or a larger parallel corpus for deriving the learned lexicon. We estimated the effects of these steps by evaluating the performance of smaller historical lexicons and a smaller parallel corpus. The smaller corpora were chosen by selecting the first 10, 100, 1000 and 10,000 sentences of the Statenvertaling bible. The smaller historical lexicons were produced by choosing the first 10, 100 and 1000 words of the words of the Statenvertaling bible sorted by decreasing frequency.

The evaluation scores are summarized in the graphs in Figure 4. The performance of the historical lexicon reaches 82% for 8,563 words (left graph). The shape of the graph suggest that increase of the lexicon size to about 60,000 would lead to a performance increase of 4%. A similar performance increase could be obtained by increasing the training text for the learned lexicon from the current 37,100 words to about 140,000 words (right graph). It will require a considerable effort to create such a large relevant sentence aligned corpus. However, the Institute of Dutch Lexicography already has compiled a historical lexicon of the required size, so there lies an interesting opportunity for further improving the quality of this approach.

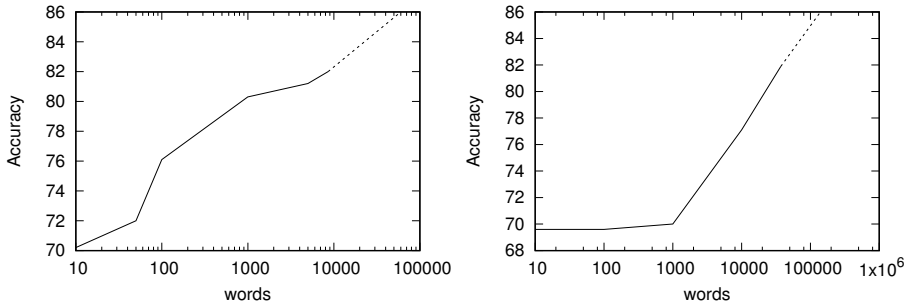

Fig. 4. POS accuracies for different lexicon sizes of the historical lexicon (left, solid line) and different corpus sizes of the learned lexicon (right, solid line). Estimated from these graphs, a historical lexicon of about 60,000 words and a parallel corpus of about 140,000 words would both lead to an increase in POS tagging accuracy from 82% to 86% (dashed lines).

5 Concluding remarks

We explored methods for improving the performance of natural language processing tools on texts written in a historical version of a language. Since the tools have been developed for processing contemporary language and languages may change considerably over time, the performance of the tools on century-old texts is usually poor. Our proposed solution is not to rebuild the tools but to translate the texts to a modern-language variety so that then they can be processed by any available tool.

We found that in order for this approach to work for the task of part-of-speech tagging, it is important that the translation was done word by word. This made it hard to employ general machine translation software for this task because it optimizes text quality by inserting, deleting and reordering words in the translated text. This makes it difficult to link annotations assigned to the modern language words back to the original historical equivalents.

Instead, we have evaluated three word-by-word translation methods for improving the quality of coarse-grained part-of-speech tag assignment to seventeenth-century Dutch text. The first was based on an online historical word to modern lemma lexicon. The second used on a lexicon that was learned from two versions of a Dutch bible, one from the seventeenth century and one from the nineteenth century. The third method employed orthographic rules learned from the learned lexicon. The rules converted historical character sequences to their modern equivalent.

We found that the two lexicon-based method performed equally well, 82% POS accuracy, where the baseline was 68% and a human translation reached 89% (see Table 1). An important difference between the two methods is that the historical lexicon method translates words to lemmas, thus eliminating the possibility of accurately assigning fine-grained POS tags. The learned lexicon does not have this disadvantage. The orthographic rules performed worse than

the two lexicon methods (73%) but they remain interesting as a possible post-processing method applied to the output of the lexicon methods.

The performance of the two lexicon-based methods is dependent on the size of the historical lexicon and the size of the parallel training texts. Based on performances with smaller lexicons and smaller training texts, we have estimated that a seven-fold increase of the size of the historical lexicon and a four-fold increase of the training corpus would both lead to a 4% improvement of POS accuracy, thus overcoming a large part of the remaining gap with the performance of human translation.

Our future work plans are all connected to extending the training and evaluation data. The available historical lexicons for Dutch are larger than the material used in this paper and we would like examine the effect of the larger lexicons on part-of-speech tagging accuracy. This study has focused on tagging texts from the seventeenth century and it would be interesting to apply these methods to other natural language processing tasks and to material of other time periods. This requires more historical material with gold standard annotation. Fortunately, today more of such data are becoming available (for example [18]).

References

[1] Archer, D., Kyto, M., Baron, A., Rayson, P.: Guidelines for normising Early Modern English corpora: Decisions and justifications. ICAME Journal 39 (2015), dOI: 10.1515/icame-2015-0001

[2] Van den Bosch, A., Busser, G., Daelemans, W., Canisius, S.: An efficient memory-based morphosyntactic tagger and parser for Dutch. In: Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, pp. 99–114. Leuven, Belgium (2007)

[3] Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., van den Bosch, A.: Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. In: Proceedings of LREC 2016. pp. 1277–1281. ELRA, Portoroz, Slovenia (2016)

[4] Fung, P., Church, K.: K-vec: A New Approach for Aligning Parallel Texts. In: Proceedings of COLING94. pp. 1096–1102. Kyoto, Japan (1994)

[5] Han, B., Baldwin, T.: Lexical normalisation of short text messages: Makn sens a# twitter. In: Proceedings of ACL HLT 2011. pp. 368–378. Association for Computational Linguistics, Portland, OR (2011)

[6] Hupkes, D., Bod, R.: POS-tagging of Historical Dutch. In: Proceedings of LREC 2016. pp. 77–82. ELRA, Portoroz, Slovenia (2016)

[7] INL: Geintegreerde Taal-Bank (GTB) (2007), instituut voor Nederlandse Lexicografie, http://gtb.inl.nl/ Retrieved 13 May 2016

[8] INL: Lexicon Service (2015), instituut voor Nederlandse Lexicografie, http://sk.taalbanknederlands.inl.nl/LexiconService/ Retrieved 13 May 2016

[9] Jiang, Jing: A literature survey on domain adaptation of statistical classifiers. http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.pdf Retrieved 13 May 2016 (2008)

[10] Kaufmann, M., Kalita, J.: Syntactic normalization of twitter messages. In: International conference on natural language processing (ICON). Kharagpur, India (2010)

[11] Koehn, P.: MOSES - Statistical Machine Translation System - User Manual and Code Guide. University of Edinburgh (2015)
[12] Koomen, N.: Van Beverwijck, Schat der Gesontheyt, 1663. http://volkoomenoude herbariaenmedisch.nl/index.beverwijck.html Retrieved 13 May 2016 (2007)
[13] McClosky, David and Charniak, Eugene and Johnson, Mark: Automatic domain adaptation for parsing. In: Proceedings NAACL HLT 2010. pp. 28–36. Association for Computational Linguistics, Los Angeles, CA (2010)
[14] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for Automatic Evaluation of Machine Translation. In: Proceedings of ACL 2002. pp. 311–318. Association for Computational Linguistics, Philadelphia PA (2002)
[15] Plank, B.: Domain Adaptation for Parsing. PhD thesis, University of Groningen, The Netherlands (2011)
[16] Rem, M., van Halteren, H.: Tagging and Lemmatization Manual for the corpus van Reenen-Mulder and the Adelheid 1.0 Tagger-Lemmatizer. Radboud University Nijmegen (2007)
[17] Reynaert, M.: Text-Induced Spelling Correction. PhD Thesis, Tilburg University (2005)
[18] Rutten, G., van der Wal, M.: Letters as Loot. A sociolinguistic approach to seventeenth- and eighteenth-century Dutch. John Benjamins (2014)
[19] Schnabel, T., Schütze, H.: Flors: Fast and simple domain adaptation for part-of-speech tagging. Transactions of the Association for Computational Linguistics (TACL) 2, 15–26 (2014)
[20] Schneider, G., Lehmann, H.M., Schneider, P.: Parsing Early and Late Modern English corpora. Digital Scholarship in the Humanities 30(3) (2015)
[21] van der Sijs, N.: Chronologisch woordenboek: De ouderdom en herkomst van onze woorden en betekenissen. Veen, Amsterdam/Antwerpen (2001)
[22] van der Sijs, N. (ed.): Biblia, dat is De gantsche H. Schrifture (Statenvertaling 1637). DBNL: Digitale Bibliotheek voor de Nederlandse Letteren (2008)
[23] Theologencommissie (ed.): Statenvertaling Jongbloededitie 1888. Statenvertaling.net Retrieved 13 May 2016 (1999)
[24] Tjong Kim Sang, E.F.: Machine Learning of Phonotactics. PhD thesis, University of Groningen, The Netherlands (1998)
[25] Van Eynde, F.: Part of Speech Tagging and Lemmatizing of the Corpus Gesproken Nederlands (Spoken Dutch Corpus). KU Leuven (2004)
[26] Wijckmans, T.: personal communication (2015)