Data-driven Morphology and Sociolinguistics for Early Modern Dutch

Marijn Schraagen **Marjo van Koppen**
Utrecht Institute of Linguistics OTS
Utrecht University
{m.p.schraagen,j.m.vankoppen}@uu.nl

Feike Dietz
Institute for Cultural Inquiry
Utrecht University
f.m.dietz@uu.nl

Abstract

The advent of Early Modern Dutch (starting ∼1550) marked significant developments in language use in the Netherlands. Examples include the loss of the case marking system, the loss of negative particles and the introduction of new vocabulary. These developments typically lead to a lot of variation both within and between language users. Linguistics research aims to characterize and account for such variation patterns. Due to sparseness of digital resources and tools, research is still dependent on traditional, qualitative analysis. This paper describes an ongoing effort to increase the amount of tools and resources, exploring two different routes: (i) modernization of historical language and (ii) adding linguistic and sociolinguistic annotations to historical language directly. This paper discusses and compares the experimental setup, and preliminary results of these two routes and provides an outlook on the envisioned linguistic and sociolinguistic research approach.

1 Introduction

In the 16th century, the language situation in the Netherlands changed substantially. One important influence is the interest in standardization of the Dutch language (Lambrecht, 1550; de Heuiter, 1581; Spiegel, 1584; Stevin, 1586). This standardization process was combined with ongoing developments in case marking (Weerman and de Wit, 1999), negation (Hoeksema, 1997), and various other lexical and morphosyntactic phenomena (Howell, 2006). The extent to which these developments were actually adopted by language users differs between and within individual language users (Bax and Streekstra, 2003; Nobels and Rutten, 2014).

1637: *Ende het gout deses lants is goet*
1888: En het goud van dit land is goed
'And the gold of that land is good'

Figure 1: Example parallel Bible translation

To date, most approaches to study these phenomena have been qualitative in nature. In this paper an ongoing effort is described to enrich Early Modern text with linguistic and sociolinguistic information in a systematic way, to allow a quantitative computational linguistic approach. The paper explores two routes to develop such an approach: a modernization route and a historical annotation route. In Section 2 an approach to text modernization is outlined. Section 3 describes an automatic tagging approach with manual post-correction and metadata enrichment. Section 4 provides a comparison between these two routes.

2 Modernization

In contrast to historical Dutch, modern Dutch is a very well-resourced language for NLP applications. A translation modernization step allows to use these resources for historical texts (Tjong Kim Sang, 2016). The modernization process can benefit from the similarity between the two historically related language varieties (Koolen et al., 2006). For the development of the modernization method described in this paper, a parallel pair of Dutch Bible translations from 1637[1] and 1888[2] is used. The close parallelism of this training pair (see Figure 1) allows for efficient application of word pair extraction algorithms. The method consists of a combination of three approaches: (i) application of manual expert rewriting rules (cf.

[1] http://dbnl.nl/tekst/_sta001stat01_01/
[2] http://www.statenvertaling.net

Braun, 2002; Robertson and Willett, 1992), (ii) extraction of a translation lexicon from parallel pairs in training data (cf. Bollmann et al., 2011), and (iii) the application of the existing statistical machine translation framework Moses (Köhn et al., 2007), cf. (Pettersson et al., 2013; Scherrer and Erjavec, 2016). In the remainder of this paper the combination of the first two approaches is referred to as the *custom method*, while the third approach is called the *SMT method*.

The dataset is split into a training part (32,235 lines, 946,721 words, 87%) and a test part (5,000 lines, 140,812 words, 13%). The split is linear, with the training part ranging from Genesis to the (apocryphic) Book of Ezra, and the test part ranging from Ezra to 3 Maccabees. For the SMT method, a subset of 2000 lines (58,249 words) is removed from the end of the training set to be used as a development set for MERT.

In Table 1 the results of the modernization approach are listed. To compare translations, the BLEU measure is used (Papineni et al., 2002). This score has notable shortcomings as a measure of translation accuracy (Callison-Burch et al., 2006), pertaining to phrase permutation and semantic unawareness. However, these shortcomings appear to be less severe for a modernization task, where phrase-based translations and word reordering are less likely to occur. Moreover, a correct translation is not the main goal of this method. Instead, the modernization is intended to increase accuracy of NLP methods, e.g., POS tagging, syntactic parsing, frequency counts, lexicon lookup, etc. It is not yet clear how well the BLEU score correlates with accuracy of these methods, however some correlation is to be expected.

The details of the custom method are as follows: as sub-baseline, **no translation** is performed. As **baseline** all parallel sentences of equal length have been extracted from the training data, and all words with an unambiguous (i.e., always the same) translation are used as a translation lexicon for the test data. Next, all sentences are **aligned** on word level to extract additional translation pairs. Note that the baseline and the alignment are relatively efficient, due to the close parallelism of the source data. Then, manual modernization **rules** are applied, specifically targeted to Early Modern Dutch, such as case marker normalization, negation normalization, clitic separation, numeral normalization. Note that phonetic rewriting is not part of this step. Next, translation pairs are constructed for **multiple word** translations (e.g., *deses* → *van dit* in Figure 1, English: this_{GEN} → of this). At this point, the test set is already sufficiently modern to allow accurate POS tagging, at least on the tokens that have been assigned the correct, modern translation. This **POS information** can be used to translate a historical word in different ways conditional on the surrounding POS tags. This is similar to the multiple word translation, except that the selection on POS tags allows to generalize over the vocabulary. As an example, the pronoun *haer* is likely to be translated as *hen* (them) before a verb, and as *hun* (their) before a noun. To limit sparseness issues, the main POS tag is used without features. For both the multiple word and the POS step, pairs have been selected using hill-climbing, implemented as incremental inclusion of those pairs that increase the BLEU score on the training set. Note that, since the pairs are extracted from the training set (i.e., a development set is not used), the hill-climbing selection is equivalent to selecting translations with a true positive application rate of over 0.5. Finally, a number of highly document-specific rules have been applied to address differences in **punctuation** between the two Bible translations. Examples of rules and word pairs are provided in Table 2.

For the evaluation of the SMT method, a different sequence of steps is applied. First, a model is built by Moses using the training set with **basic** settings. Then, **MERT tuning** is applied using the development set. Next, the **capitalization** model of Moses, which turned out to be highly inaccu-

method	*steps*	BLEU
custom	no translation	0.134
	baseline	0.507
	aligned	0.530
	rules	0.581
	multiple word	0.600
	POS information	0.619
	punctuation	0.627
SMT	Moses basic	0.597
	MERT tuning	0.616
	capitalization	0.639
combination	rules	0.644
	multiple word, POS	0.647
	punctuation	0.653

Table 1: Translation evaluation

input	output	notes	translation
rewriting rules			
stem+se	*stem* ␣ hen	pronoun clisis	them
eens *stem*s	van een	genitive	of a
den/mijnen/welken	de/mijn/welke	case loss	the/mine/which
alle de	al de	agreement loss	all the
en [. . .] *negative*	*negative*	negative concord	
numeral ␣ ende ␣ *num*	*num*+en+*num*		
num ␣ *num*	*num*+*num*		
punctuation rules			
`␣	␣		
; *[upper case]*	:		
; *[lower case]*	,		
said, *[upper case]*	:		
extracted multiword pairs			
haer zelfden	zichzelf	reflexive pronoun	him/her/it/them/oneself
waer heen	waarheen	prepositional compound	where to
leeuws tanden	leeuwentanden	case loss, compounding	lion teeth
potte-backers kruik	pottenbakkerskruik	case loss, compounding	pottery jar
rechteroog heupe	rechterheup	terminology shift	right hand side hip
extracted Part-of-Speech pairs			
alle+V	allen		all
alle+PRON	al		all
PUNCT+alle	al		all
daar+V	er		there

Table 2: Example implementations of translation steps

rate, is corrected using post-processing. Combining the SMT and the custom method, manual rewriting **rules** are applied on top of SMT, followed by **multiple word** alignment, **POS** information and **punctuation rules**.

2.1 Discussion

Both the method using manual rules combined with automatic translation pair extraction as well as the method using the Moses toolkit show a substantial improvement over the baseline performance. For the first method, the manual rule component provides the largest share of the performance improvement. This indicates (consistent with, e.g., Pettersson et al., 2012) that language development over time, at least in the case of Bible translations, displays a high level of regularity, which can be captured by a small number of rewriting rules. Interestingly, the morphological rules combined with translation pair extraction offer sufficient coverage to omit phonetic rewriting commonly used in language modernization. Note that this behavior depends on similarity between training and test vocabulary, which will be discussed further in Section 2.1.1.

The described method provides competitive performance as compared to the SMT approach. It can be considered promising that the results of a state-of-the-art machine learning algorithm can be reproduced using a relatively straightforward approach. However, Table 1 also shows that the combination of approaches offers very little improvement over the performance of the SMT algorithm in isolation. Therefore, it is at present not fully clear how to incorporate the custom translation pair extraction or manual morphological rules into a combined methodology.

To obtain a better insight in the performance of the various methods, a more extensive evaluation is necessary. This includes the application of the method on more diverse data and a systematic comparison between approaches. Furthermore, the evaluation could be extended into a more application–oriented direction, i.e., by analyzing results of NLP methods on modernized text.

2.1.1 CLIN27 Shared Task

The method presented in this paper has also been entered into the CLIN27 Shared Task on Translating Historical Text[3]. The results of the system on this task are considerably lower than in the present evaluation, which can be contributed to several factors.

First, the test set used for the Shared Task was markedly different from the provided training set. The test set contained genres such as theater plays, letters, eulogies, administrative texts, journal entries, and bullet point lists of activities. These genres introduce a significant amount of new vocabulary, for which the word-level vocabulary-based method as presented in this paper is not particularly well-suited. In the Shared Task the method was extended with a set of phonetic rewriting rules, which showed a large performance increase. This is consistent with previous work on character-level SMT approaches (e.g., Scherrer and Erjavec, 2013), which are essentially a way to automatically extract phonetic rewriting rules from data.

Furthermore, the test set contained texts ranging from 1607 to 1692. Various morphosyntactic or spelling-related phenomena occurring in the 1637 training set which are targeted with specific rules do not occur in later texts, such as negative concord constructions. Application of these rules on later texts actually decreases performance in certain cases, and should therefore be controlled by time period constraints.

Additionally, the test set fot the Shared Task was created with the specific goal of word-level spelling modernization to facilitate POS tagging (cf. Tjong Kim Sang, 2016). This resulted in a rather artificial translation, preserving sentence length and word order, leaving historical word forms untranslated in case a modern tagger already assigned a correct tag. As a result, in several cases the current method provides an arguably better translation which is nevertheless evaluated as an error. Further analysis showed that, for a number of participants in the Shared Task, manually respelling a very small number of frequent errors resulted in a substantial performance improvement.

For the reasons mentioned above, the results on the CLIN27 Shared Task should not be considered as a conclusive evaluation of the current method. However, the results do indicate important aspects of the current method, such as the impact of training vocabulary, and the influence of the goal application on the translation requirements. Further development of these issues is ongoing in the current project.

[3] https://ifarm.nl/clin2017st/

3 Annotation

As stated above, text modernization allows for the use of resources and tools for contemporary language. However, this approach also introduces incorrectly translated and non-translated tokens, which limit the accuracy of NLP applications. Moreover, certain information from the historical text is lost. Modernization entails spelling normalization, which means that, e.g., spelling differences over time can no longer be studied. Other topics of research, such as case marking or negation, may also be lost after modernization, or it may prove difficult to link the modernized text to the historical original. Therefore, an additional research direction processes historical text directly, using tools and resources for historical language or using manual annotation. The annotation effort also allows extension to sociolinguistic information, which is intrinsically outside the scope of modernization approaches. The remainder of this section describes the setup of the annotation task, which are currently ongoing.

3.1 Part-of-speech tagging

A pilot project has been designed to annotate a corpus of letters by the Dutch author and politician P.C. Hooft, written between 1600 and 1647. In the absence of tools for Early Modern Dutch, a POS tagger for Middle Dutch (1200–1500) is used (van Halteren and Rem, 2013). Although Middle Dutch is considerably different from Early Modern Dutch, several properties of interest are shared, such as case marking and negation clitics. Therefore, a tagger capable of marking such properties is preferred over contemporary equivalents. However, as expected on Early Modern data, the overall accuracy of the tagger is low. Therefore, a manual annotation effort is ongoing to check and correct all assigned tags (including morphosyntactic features) in the corpus manually.

3.2 Sociolinguistic tagging

An accurately tagged corpus allows to discover patterns on a morphosyntactic level. To analyze the development of such phenomena, non-linguistic variables have to be taken into account.

Als	*nu*	*de*	*veerschujt*	*niet*	*anders*	*ujt*	***en***	*leverde*	*dan*	*den*	*brief*
Con(sub)	Adv(gen)	Det()	N(sg)	Pro(neg)	Adj(-s)	Adp(prtcl)	**Adj(negcl)**	V(past)	Con(cmp)	Det(-n)	N(sg)
If	now	the	ferry	nothing	else	out	**not**	delivered	than	the	letter

'If the ferry would not deliver anything but the letter'

Figure 2: Example tagged sentence, showing a negation clitic

The letter corpus contains dated documents, therefore a straighforward variable is time, allowing analysis of when certain developments have occurred. Other variables of interest include the topic of correspondence, the type of relation between the correspondents and the domain of the correspondence (government, finance, literature, etc.), the goal of the correspondence (invitation, recommendation, request, etc.), and personal information about the correspondent (age, gender, literary status). Furthermore, the rhetorical structure of a text is annotated, in terms of greeting, opening, body, closing. This for instance allows to verify the hypothesis that certain parts of letters, e.g., the opening and closing sections, are highly formulaic, and therefore do not exhibit language development to the same degree as the body text (Nobels and Rutten, 2014). Annotation is performed by a pool of nine annotators. To measure inter-annotator agreement, 10% of the corpus is assigned to random pairs of annotators. The full list of sociolinguistic variables is provided in the Appendix.

In Figure 2 an example sentence with part-of-speech tags is provided. This sentence contains the negation clitic *en*, alongside the main negation *niet*. Once the full corpus is properly tagged this clitic can be studied systematically, e.g., to investigate the neighbouring tags or lemmas of the clitic, or to check whether or not the clitic is used more often in formal writing.

3.3 Interoperability

To increase the practical accessibility of the annotation data, a collaboration with the Nederlab project (Brugman et al., 2016) has been established. Nederlab provides an online search interface for the data in the Digital Library of Dutch Literature[4] using Corpus Query Processor (Evert and Hardie, 2011), which allows to search for linguistic annotation and metadata. For this collaboration, several interoperability issues need to be addressed. The Adelheid tagger uses the CRM tagset, which contains a set of features specific for Middle Dutch. The Nederlab project uses the CGN tagset (van Eynde et al., 2000), for which both the main tags and the feature set differ considerably from CRM. For the current pilot several additional features are introduced to facilitate the analysis of language development.

Apart from the tagset, the output format needs to be converted as well. Nederlab uses the FoLiA format (van Gompel and Reynaert, 2013), which is a de facto standard XML linguistic annotation format for Dutch, whereas Adelheid uses a custom XML format. To facilitate integration with current annotations and metadata in Nederlab, a word-level alignment of the FoliA output is planned.

Further interoperability considerations include incorporation of linked data, e.g., for correspondents in the current dataset which may also be found in encyclopedic resources, and using existing classification schemes, such as HISCO for historical occupational titles (van Leeuwen et al., 2002).

4 Data-driven historical linguistics

The two methods outlined in this paper are intended to complement eachother in providing an environment for computational historical linguistics research. Modernization has the advantage that research questions can be addressed using the existing infrastructure for a modern language, in terms of resources, approaches, evaluation data et cetera. The disadvantage of this method is the inherent loss of information and the occurence of translation errors, which entails that several topics of interest cannot be studied using modernized data, or that the validity of results is unclear. In contrast, manual enrichment provides high-quality linguistic annotations as well as the possibility to include meta-linguistic information. The obvious disadvantage of this method is the large amount of time and/or financial resources necessary. However, if a sufficiently large amount of data is an-

[4] `http://www.dbnl.org`, in Dutch

notated (possibly in combination with automatically derived annotations, cf. Hupkes and Bod, 2016), machine learning algorithms can be trained to allow for automatic annotation. The combination of modernization and manual annotation may prove valuable as a methodology in historical (socio-)linguistics. Future work in the current project, however, is necessary to validate this claim.

References

Marcel Bax and Nanne Streekstra. 2003. Civil rites: ritual politeness in early modern Dutch letter–writing. *Journal of Historical Pragmatics*, 4(2):303–325.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop*, pages 34–42. ACL.

Loes Braun. 2002. Information retrieval from Dutch historical corpora. Master's thesis, Maastricht University.

Hennie Brugman, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, and Antal van den Bosch. 2016. Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In *Proceedings of LREC 2016*.

Chris Callison-Burch, Miles Osborne, and Philipp Köhn. 2006. Re-evaluation the role of BLEU in machine translation research. In *Proceedings of EACL*, pages 249–256. ACL.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham.

Frank van Eynde, Jakub Zavrel, and Walter Daelemans. 2000. Part of speech tagging and lemmatisation for the Spoken Dutch Corpus. In *Proceedings of LREC 2000*.

Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML format for linguistic annotation-a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81.

Hans van Halteren and Margit Rem. 2013. Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters. *Language Resources and Evaluation*, 47(4):1233–1259.

Pontus de Heuiter. 1581. *Nederduitse orthographie*. Edited by G.R.W. Dibbets, 1972, Wolters-Noordhoff.

Jack Hoeksema. 1997. Negation and negative concord in Middle Dutch. *Amsterdam Studies in the Theory and History of Linguistic Science*, 4:139–156.

Robert Howell. 2006. Immigration and koineisation: the formation of Early Modern Dutch urban vernaculars. *Transactions of the Philological Society*, 104(2):207–227.

Dieuwke Hupkes and Rens Bod. 2016. Pos-tagging of historical Dutch. In *Proceedings of LREC 2016*.

Philipp Köhn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Marijn Koolen, Frans Adriaans, Jaap Kamps, and Maarten de Rijke. 2006. A cross-language approach to historic document retrieval. In *ECIR 2006: Proceedings of the 28th European Conference on IR Research*, pages 407–419. Springer.

Joos Lambrecht. 1550. *Nederlandsche spellijnghe, uutghesteld by vraghe ende antwoorde*. Edited by J.F.J. Heremans and F. Vanderhaeghen, 1882, C. Annoot-Braeckman.

Marco van Leeuwen, Ineke Maas, and Andrew Miles. 2002. *HISCO: Historical International Standard Classification of Occupations*. Cornell University Press.

Judith Nobels and Gijsbert Rutten. 2014. Language norms and language use in seventeenth-century Dutch: negation and the genitive. In Gijsbert Rutten, editor, *Norms and usage in language history, 1600-1900. A sociolinguistic and comparative perspective.*, pages 21–48. John Benjamins Publishing Company.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2012. Rule-based normalisation of historical text - a diachronic study. In *Proceedings of the First international Workshop on Language Technology for Historical Text*, KONVENS, pages 333–341.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics at NODALIDA*, pages 54–69. Linköping.

Alexander Robertson and Peter Willett. 1992. Searching for historical word-forms in a database of 17th-century English text using spelling-correction methods. In *Proceedings of ACM SIGIR '92*, pages 256–265. ACM.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based SMT. In *BSNLP 2013-4th Biennial Workshop on Balto-Slavic Natural Language Processing*.

Yves Scherrer and Tomaž Erjavec. 2016. Modernising historical slovene words. *Natural Language Engineering*, 22(6):881–905.

Hendrik Spiegel. 1584. *Twe-spraack. Ruygh-bewerp. Kort begrip. Rederijck-kunst*. Edited by W.J.H. Caron, 1962, Wolters-Noordhoff.

Simon Stevin. 1586. *Uytspraeck van de weerdicheyt der Duytsche tael*. Chr. Plantijn.

Erik Tjong Kim Sang. 2016. Improving part-of-speech tagging of historical text by first translating to modern text. In *Proceedings of the International Workshop on Computational History and Data-Driven Humanities*, pages 54–64. Springer.

Fred Weerman and Petra de Wit. 1999. The decline of the genitive in Dutch. *Linguistics*, 37(6):1155–1192.

Appendix: sociolinguistic variables

- Purpose of the letter
  - Express thanks
  - Compliment/praise
  - Excuse
  - Ask for a favour
  - Ask for information
  - Ask for advice
  - Admonish
  - Inform
  - Remember
  - Persuade
  - Order
  - Allow
  - Invite
- Topic of the letter
  - Business
  - Literature
  - Domestic affairs
  - Love
  - Death
  - News
  - Religion/ethics
- Correspondent information
  - name
  - group or individual
    for individuals:
  - birth/death date
  - gender
  - occupation
  - literary author
  - relation to P.C. Hooft
- Letter structure
  - Introductory greeting
  - Opening (optional)
  - Narratio
  - Closing (optional)
  - Final greeting