An automatic part-of-speech tagger for Middle Low German

Mariya Koleva, Melissa Farasyn, Bart Desmet, Anne Breitbarth and Véronique Hoste
Ghent University

Syntactically annotated corpora are highly important for enabling large-scale diachronic and diatopic language research. Such corpora have recently been developed for a variety of historical languages, or are still under development. One of those under development is the fully tagged and parsed Corpus of Historical Low German (CHLG), which is aimed at facilitating research into the highly under-researched diachronic syntax of Low German. The present paper reports on a crucial step in creating the corpus, viz. the creation of a part-of-speech tagger for Middle Low German (MLG). Having been transmitted in several non-standardised written varieties, MLG poses a challenge to standard POS taggers, which usually rely on normalized spelling. We outline the major issues faced in the creation of the tagger and present our solutions to them.

Keywords: historical linguistics, part-of-speech tagging, conditional random fields, feature selection, normalization

1. Introduction

Corpora of historical texts annotated with different levels of grammatical information, such as parts of speech, (inflectional) morphology, syntactic chunks, clausal syntax, provide an important resource for studies of diachronic syntactic variation and change (e.g. Kroch et al. 2000, Rögnvaldsson & Helgadóttir 2011). They enable the automatic extraction of syntactic information from historical texts (more than is manually possible), and allow making statistically valid observations. Apart from reducing the amount of time required for data retrieval, an important advantage is that they make research testable and replicable. The Corpus of Historical Low German (CHLG)

(Breitbarth et al. 2011, 2012) is in the process of becoming a corpus of syntactically parsed texts of Old Low German (OLG)/Old Saxon and Middle Low German (MLG). The corpus will facilitate – currently still urgently lacking – research into the diachronic syntax of Low German, a language that is geographically and linguistically located between High German in the south(east) and Dutch and Frisian in the north(west), and hence still forms a missing link within Continental West Germanic (CWG). Eventually, the CHLG Corpus will fill the recently growing ranks of a family of parsed historical corpora including the Penn Parsed Corpora of Historical English (Mitchell et al. 1993), the Tycho Brahe Parsed Corpus of Historical Portuguese (Brito et al. 2002), the parsed corpus of historical French Modéliser le Changement: Les Voies du Français (MCVF) (Martineau 2005) and the Icelandic Parsed Historical Corpus (IcePAHC) (Wallenberg et al. 2011).

The OLG stage being completed (Walkden 2016), the work is currently focusing on the MLG stage. As the corpus is intended to serve as an important resource for the study of syntactic variation and change, it will be annotated with syntactic information. The first step in this syntactic enrichment of the corpus consists in the assignment of parts-of-speech tags to every word in the corpus. Tools to automatically assign high-accuracy POS tags are freely available for a variety of languages, but they rely on large amounts of annotated training data. Those that have been trained on contemporary varieties of text tend to suffer a drop in performance when applied to older varieties and adapting them for use on historical language remains challenging (Moon & Baldridge 2007, Bennett et al. 2010). We adopt a data-driven approach to POS tagging in which we try to overcome the issues of small datasets with considerable corpus-internal diatopic, diachronic, stylistic, and spelling variation, and pay particular attention to the choice of rich linguistic features, the choice of a robust machine learning algorithm, and the potential benefit of using genetic algorithms to optimize the feature space.

In this paper, we report on the work leading to the creation of such an optimized part-of-speech tagger for MLG. We present the current design of the corpus and give an overview of the language to offer an idea of the difficulties one has to deal with when creating a corpus of an under-researched, naturally non-standard, historical language. We compare our own way of working to a background of methods and standards from related work on corpora concerning the same period or geographical area. Subsequently,

we present the experiments and results of the research on the creation of the automatic POS tagger. We also describe the role of normalization and of fine-grained inflectional morphology in potentially improving the tagger's accuracy.

In Section 2, we outline related work that has been carried out in projects similar to CHLG with a special view on the development of POS taggers. Section 3 will present the current set-up and aims of the CHLG corpus, highlighting the premises of the current text selection and the subsets thereof used for work on the tagger. Section 4 presents the methodology and the set-up of the experiments, the results of which are discussed in Section 5. Section 6 addresses further experiments showing how spelling normalization and the addition of morphological information can further improve the accuracy of the tagger. Section 7 concludes this paper.

2. Related work on historical corpora of German

Building a historical corpus is a task that essentially involves dealing with the same difficulties as building a synchronic corpus, such as issues of size, representativeness, sampling, etc. (Biber et al. 1998, Bennett et al. 2010). In contrast to synchronic corpora, however, diachronic corpora are restricted by the character of the written attestation of the language, which is typically limited as regards the amount and genres of text transmitted. Furthermore, they often lack a standardized orthography, complicating in particular automated processing such as POS tagging. In order to provide a background for the discussion of creating a POS tagger for the CHLG, the current section discusses previous work on similar corpora of historical varieties of (Low) German and other languages, and the computational tools associated with them.

2.1 Other varieties of historical Low German

Linguistic and particularly syntactic research on Middle Low German has been limited due to a lack of digitized data that can be searched easily. While there have been several efforts to create corpora of Middle Low German, none of them involved syntactic

annotation. The Atlas spätmittelalterlicher Schreibsprachen des niederdeutschen Altlandes und angrenzender Gebiete (AsnA) (Peters & Fischer 2007) and the Atlas Spätmittelalterlicher Schreibsprachen des ostniederdeutschen Raumes (AsoR) (Bieberstedt 2015) had as their main goal to create a resource enabling research into the variation between the MLG scribal languages over time, especially at the level of spelling and semantics. The Atlas is therefore not readily suited for syntactic research, aside from the fact that is not yet publically available. The project Textverdichtungsprozesse im Spätmittelalter (Textual compacting processes/complexification in the late Middle Ages) at Paderborn University (Tophinke 2009, 2012; Tophinke & Wallmeier 2011) explored, first transcribed and partially annotated MLG legal texts to trace the formal development of linguistic complexification in MLG, as well as its dynamics, speed, and regional spread. The project, however, has a narrow focus on legal texts.

The only project comparable to the CHLG in aims and methods (covering the diachronic, diatopic and stylistic variation of MLG), except for the syntactic annotation, is the Referenzkorpus Mittelniederdeutsch und Niederrheinisch (ReN) (Peters & Nagel 2014, Schröder 2014). The aim of the ReN is the creation of a corpus of transcribed, lemmatized and grammatically annotated texts from 1200 until 1650. The grammatical annotation consists of POS and inflectional morphology, but not syntactic parsing as is the aim for the CHLG. The POS and morphological tagging of the ReN was assisted by *RFtagger* (Schmid & Laws 2008), built into the *CorA* annotation tool, and manually corrected by two annotators. While distinguishing itself by its speed, *RFtagger* has the disadvantage of being inflexible when it comes to adapting parameters for optimization based on features of the texts. Its current accuracy for POS tagging MLG texts is between 83.9% and 90% in-domain (trained on and applied to the same text) and between 69.3% and 73.5% out-of-domain, depending on the data set (Barteld et al. 2015). The CHLG works closely together with the ReN for the POS and morphological annotation of the texts contained in both corpora, but uses more adaptive automated taggers, as reported in more detail in Section 4 below.

2.2 Related language varieties of the same period and geographical area

In contrast to MLG, for which digitized data is scarce, and linguistically annotated corpora even more so, different corpora have already been built for historical varieties of High German, such as the Referenzkorpus Altdeutsch (www.deutschdiachrondigital.de), the Referenzkorpus Mittelhochdeutsch (Dipper 2015) (http://referenzkorpus-mhd.uni-bonn.de), the Bonn corpus of Early New High German or Bonner Frühneuhochdeutschkorpus (Diel et al. 2002, Fisseni et al. 2007) (https://korpora.zim.uni-duisburg-essen.de/Fnhd), the GerManC corpus (Scheible et al. 2011a, 2011b) (www.llc.manchester.ac.uk/research/projects/germanc) and the German Text Archive or DTA (Geyken et al. 2011) (www.deutschestextarchiv.de).

The annotated Referenzkorpus Altdeutsch covers older varieties of German, viz. Old High German and Old Saxon (Old Low German), and aims to cover all currently surviving texts in Old German from 750 to 1050 AD. It comprises around 650,000 words from texts (mostly from editions) digitized within the TITUS project. The largest subset of the corpus consists of translations from Latin and mixed German-Latin texts (Linde & Mittmann 2013), primarily religious ones. The Referenzkorpus Altdeutsch was semi-automatically pre-annotated with information extracted from grammars and glossaries. The layers of annotation involve metadata (e.g. genre, origin, period), structural information (lines, paragraphs, sentences, words, etc.), POS, inflectional morphology, and syntactic information. All word forms were lemmatized. Some POS tags could be automatically extracted by comparing tokens to a list with existing glossaries and some morphological features belonging to them (e.g. conjunctions). The inflectional information was extracted partly automatically, but only if the grammar provided enough information, for instance for the strong verb classes.

The Bonner Frühneuhochdeutschkorpus covers the period from 1350 to 1700 with one (relatively short, ca. 400 words) text from each of the seven fifty-year time slots and ten regional varieties (covering Upper and Central German). The only POS tags (in the "typ" attribute to the "wortform" tag) are "substantive", "adjektiv", "verb", "unbekannt" (unknown) or "potentiell" (for contextually undecidable forms). Besides, there is morphological information (Diel et al. 2002). The GerManC corpus covers the period from 1650 to 1800 in 50-year increments and includes five regional dialects (North, West Central, East Central, West Upper and East Upper German). Its size is

about 1,000,000 tokens with the sampling done by taking three extracts of 2,000 words per genre, period and region. Gold standard annotations for lemmas, POS tags and normalized spelling were added to a representative subset of the corpus. The DTA covers a comparable period of Early Modern German (1650 - 1900). The selected texts were automatically tokenized, lemmatized, and POS tagged.

Choosing a POS annotation scheme is crucial for any annotation project, but this especially holds for diachronic corpus projects, because of the dynamicity especially in word classes prone to grammaticalisation (e.g. prepositions, non-finite verb forms and pronouns). In the case of GerManC, a well-established tagset for Modern German – the Stuttgart-Tübingen Tagset (STTS) (Schiller et al. 1995) – was adapted to reflect specific phenomena in Early Modern German. Six new tags were added to the tagset. Overall, the inter-annotator agreement on the POS task was 91.6%. (Scheible et al. 2011a, 2011b). The POS annotation in the Referenzkorpus Altdeutsch on the other hand uses the DDDTS tagset (Deutsch Diachron Digital tagset), which is likewise based on STTS. Similar to HiTS (Dipper et al. 2013), it uses two-level annotation (the lemma and the concrete realisation as instantiated by the word form) so as to capture language change.

3. The Corpus of Historical Low German

The Corpus of Historical Low German was conceived to foster linguistic and particularly syntactic research on Middle Low German. In this section, we provide some background on the language (Section 3.1), specify the purpose of the corpus (Section 3.2) and present the corpus design (Section 3.3).

3.1 Middle Low German

Middle Low German (MLG) was the language spoken in northern Germany from about 1150 until 1600 and replaced Latin as the written language from about 1300 onwards (Peters 2003). Between 1550 and 1650, its position as the written language in the area gradually waned in favour of Early New High German. The language, however, still

exists as a spoken language (Modern Low German/Plattdeutsch), i.e. as a group of dialects without full standardization. Nevertheless, the MLG period saw a certain amount of levelling of regional variation. During the 14th century, several regional scribal languages/regionale Schreibsprachen emerged (Peters 1973, 2003; Sanders 1982), which incorporated characteristics from the surrounding dialects. MLG gained importance as the international lingua franca around the North and Baltic Seas in the 14th and 15th centuries, in connection to the expansion of the Hanseatic League of trade. Through this expansion, it also noticeably influenced several other languages, in particular the mainland Scandinavian languages (Braunmüller 1996, 2002).

3.2 Purpose of the corpus

One of the most important issues in corpus construction is to determine the purpose for which it is going to be used. The main purpose of the CHLG is to enable research of diachronic and geographical variation in historical Low German syntax and morphology. Therefore, for the MLG part of the corpus, texts from all the main scribal languages, covering the whole period of attestation (c. 1300 – 1600), and only texts that are dated and localised will be included in the corpus (Section 2.3). For the text selection, as well as the POS and morphological tagset (see Section 3.3 below), there is a close collaboration with the Referenzkorpus Mittelniederdeutsch und Niederrheinisch (ReN) mentioned in Section 2.1 (Peters & Nagel 2014, Schröder 2014). Part of the texts (both in ReN and CHLG) were transcribed within the ASnA project, while others come from the project Niederdeutsch in Westfalen, and some were newly transcribed. While the ReN only annotates for POS and morphology, the CHLG is currently adding syntactic parsing, following the Penn Treebank system to ensure interoperability with related corpora such as the PPCHE or the IcePAHC.

3.3 Corpus design

In order to enable geographical and diachronic research, it is desirable that all the texts in the corpus be dated and localized. So far, we included texts belonging to three different text types meeting this condition: charters, legal texts, and (medical and religious) prose. This makes the corpus balanced concerning scribal language as well as concerning genre. The texts also have to cover the whole period in which MLG was written, that is c. 1300 - c. 1600. Thanks to the economic and political importance of MLG in this period, many texts were produced and have survived, making it possible to cover MLG attestation in the corpus without significant gaps. Furthermore, CHLG includes texts from the three main scribal languages of MLG: Westphalian, Eastphalian and North Low Saxon. Five data point clusters spread out throughout the language area were selected, three from the Altland (west of the Elbe) and two from the Neuland (east of the Elbe). Texts from Münster, Herford, Rüthen and Soest represent Westphalian in the Altland. Braunschweig (Eastphalian, Altland), Magdeburg (Elbe Eastphalian, Neuland), Oldenburg (North Low Saxon, Altland) and Lübeck (North Low Saxon, Neuland) complete the corpus. The CHLG hence only covers texts from the present-day German area; MLG texts localized outside this area, e.g. in the mainland Scandinavian language area or the area of the eastern Baltic Sea, are not included, because of the even greater distance (than in the core MLG language area) between the written MLG and the languages spoken in these areas. Table 1 enumerates details about the texts that are included in the corpus so far. The total number of words will be around 722,000. For the POS experiments reported in the current paper, only Westphalian texts were used (55,777 words in total), as discussed in Section 4. The reason for this restriction is that the research reported here is embedded into a project on domain extension.

Table 1. Details of current texts selected for inclusion in the CHLG

Place	**Scribal language**	**Genre**	**Period/year**	**Name**	**Number of tokens**
Soest	Westphalian (Altland)	legal texts	c. 1367	*Soester Schrae*	8,241
Herford	Westphalian (Altland)	legal texts	1375	*Herforder Rechtsbuch*	16,227
Rüthen	Westphalian (Altland)	legal texts	3 parts: c. 1300, c. 1350, 1460-1500	*Statuarrecht Rüthen*	6,804
Münster	Westphalian (Altland)	religious prose	1444	*Spieghel der leyen*	24,505
Münster	Westphalian	religious	1480	*Dat myrren*	ca. 91,000

	(Altland)	prose		*bundeken*	
Münster	Westphalian (Altland)	charters	14th - 15th c.	*Urkundenbuch Münster*	ca. 95,500
Oldenburg	North Low Saxon (Altland)	charters	14th - 15th c.	*Oldenburger Urkunden*	28,241
Oldenburg	North Low Saxon (Altland)	legal texts	1336	*Oldenburger Sachsenspiegel*	24,377
Lübeck	North Low Saxon (Neuland)	charters	13th - 15th c.	*Urkundenbuch Lübeck*	ca. 179,000
Braunschweig	Eastphalian (Altland)	charters	13th - 15th c.	*Urkundenbuch Braunschweig*	ca. 81,000
Magdeburg	Elbe-Eastphalian (Neuland)	charters	13th - 15th c.	*Magdeburger Urkundenbuch*	ca. 39,000
Magdeburg	Elbe-Eastphalian (Neuland)	medical prose	1483	*Promptuarium medicinae*	ca. 128,000
Total					**ca. 722,000**

Summing up, other corpora of historical varieties of (Low) German and related languages cannot readily be compared to the CHLG concerning the computational tools used. Only the ReN, with which the CHLG is collaborating, uses similar tools and methods. The main advantage of the CHLG over the ReN approach to POS tagging, as we will show in the next section, is that we use a machine learning approach with a customisable feature space. This allows us to make full use of the bare data for training. It means that, in comparison to projects such as Referenzkorpus Althochdeutsch, we have more freedom, as there is no need to compare to existing lists of word/grammatical phenomena. Moreover, it allows us to make use of the detailed transcription and thus to circumvent issues of non-standard spelling.

4. Methodology

In order to find an efficient way to produce a POS enriched corpus of MLG, we experimented with different part-of-speech taggers to automatically assign POS information to the word forms in the corpus and evaluated their performance both with respect to accuracy and robustness. In this section, we motivate the choice of tagset, discuss the different types of experiments we set up and specify the experimental data on which the taggers were trained and tested.

4.1 Tagset

We used an existing tagset specifically designed for historical German rather than trying to derive one from the data by bootstrapping/unsupervised learning as done in particular for MLG by Sukhareva & Chiarcos (2016), as our goal was not development of a new tagset, but of a POS tagger with as high an accuracy as possible.

For the annotation, we made use of the *CorA* annotation tool (Bollmann et al. 2014), a web-based annotation tool specifically developed for tagging historical and non-standard languages. We used an adaptation of the above-mentioned HiTS tagset (Historisches Tagset) (Dipper et al. 2013), the HiNTS (Historisches Niederdeutsches Tagset) developed specifically for MLG by the Referenzkorpus Mittelniederdeutsch/Niederrheinisch (https://vs1.corpora.uni-hamburg.de/ren/). Table 2 illustrates the mapping between STTS, from which HiTS is derived, HiTS, and HiNTS. Being derived from HiTS, HiNTS is rather close to HiTS, but in case of any differences, HiNTS tends to be more fine-grained and suited to the MLG situation.

Table 2. Mapping between STTS for German, HiTS for historical German and HiNTS for MLG

STTS tags	**HiTS tags**	**HiNTS tags**
ADJA, ADJD	ADJA, ADJD, ADJN, ADJS	ADJA, ADJD, ADJN, ADJS, ADJV, ADJA<VVPS, ADJD<VVPP, ADJV<VVPP
ADV, PWAV	AVD, AVG, AVNEG, AVW	AVD, AVNEG, AVREL, AVW, AVKO
APPR, APPRART, APPO	APPO, APPR	APPO, APPR
CARD	CARDA, CARDD, CARDN, CARDS	CARDA, CARDD, CARDN, CARDS
ART, PDAT, PDS, PIAT, PIDAT, PPER, PRF	DDA, DDART, DDD, DDN, DDS, DIA, DIART, DID, DIN, NIS, PPER, PRF	DDA, DDART, DDN, DPDS, DIA, DIART, DID, DIN, DPIS, DNEGA, DNEGD, DNEGN, DPNEGS, PPER, PRF
DRELS, PRELAT, PRELS	DRELS	DRELA, DRELN, DPRELS,
PPOSAT, PPOSS	DPOSA, DPOSD,	DPOSA, DPOSD, DPOSN,

	DPOSGEN, DPOSN, DPOSS	DPOSS
PWAT, PWS	DWA, DWD, DWN, DWS	DWA, DWD, DWN, DPWS
FM	FM	FM
ITJ	ITJ	ITJ
KON, KOUS, KOKOM, KOUI	KO*, KON, KOUS, KOKOM	KO*, KON, KOUS, KOKOM, KOUI
NN, NE	NA, NE	NA, NE, NA<VVINF
PTKA, PTK, PTKANT, PTKNEG, PTKVZ, PTKZU, APZR, TRUNC	PTKA, PTKANT, PTKINT, PTKNEG, PTKREL, PTKVZ	PTKA, PTKN, PTKANT, PTKNEG, PTKVZ, PTKZU
PAV, PWAV	PAVAP, PAVD, PAVG, PAVREL, PAVW	PAVAP, PAVD, PAVREL, PAVW, PAVKO, PAVKON
VAFIN, VAIMP, VAINF, VAPP, VVFIN, VVIMP, VVINF, VVIZU, VVPP, VMFIN, VMINF, VMPP	VAFIN, VAIMP, VAINF, VAPP, VAPS, VVFIN, VVIMP, VVINF, VVPP, VVPS, VMFIN, VMIMP, VNINF, VMPP, VMPS	VAFIN, VAIMP, VAINF, VAPP, VAPS, VVFIN, VVIMP, VVINF, VVPP, VVPS, VMFIN, VMIMP, VNINF, VMPP, VMPS

To assess the quality of the tagset, two annotators with a solid background in historical linguistics both annotated independently a dataset of 3,003 tokens from Soester Schrae and 3,005 for Statuarrecht Rüthen. The former enumerates the city's laws and the latter describes the statutes of the inhabitants of Rüthen. Though from different cities, both texts belong to the same scribal language (Westphalian). After that, the inter-annotator agreement (IAA) was measured, the errors discussed and some agreements on doubtful tags were created. The pairwise average IAA for Soester Schrae was 91.4% or 0.908 Fleiss Kappa. The corresponding scores for Statuarrecht Rüthen were 92.2% and 0.924. Both the high accuracy scores and the Fleiss Kappa scores thus show an almost perfect agreement.

4.2 Standard experimental set-up

It is crucial for the success of a part-of-speech tagger that it has access to highly informative textual information in order to obtain good performance. This textual information typically consists of specific information on the linguistic unit under consideration (in our case, a token) and its manually annotated class (part-of-speech tag). This information is presented to the machine learner (discussed in Section 4.3) in a fixed format, so-called feature vectors, which ideally contain all possible disambiguating information to allow the learner to differentiate between different POS classes. We relied on two types of features, viz. so-called standard features which are typically used in the construction of POS taggers and a set of custom features, which seek to describe the specificity of the XXX data. The following standard features were implemented, an example of which is shown in Table 3:

i. Is the first letter capitalized? (binary feature);
ii. Does it contain a digit? (binary feature);
iii. Is it punctuation? (binary feature);
iv. Is it hyphenated? (binary feature which checks for hyphens and dashes, but also for markers of separate and compound writing, e.g., ghe#delet);
v. Word length (numeric feature);
vi. First n letters (bigram, trigram);
vii. Last n letters (bigram, trigram);
viii. Lowercase form of the token (symbolic feature).

In addition to this, the POS tagger also relied on three corpus-specific features:

i. A series of binary paratext features which check for notes, corrections, and expansions of certain collapsed forms. More specifically, these features encode whether the token includes editorial marks of the following types: interlinear notes or corrections, editorial additions, notes or corrections in one of the margins, interlinear expansions or expansions in one of the margins, abbreviation expansions, strikethrough text, etc. For instance, in p{R_er}sonen (persons), R_ indicates that an abbreviated form has been expanded;.

ii. A binary feature which checks whether the token is Vortmer. This token frequently marks the beginning of a new sentence or section, even in the absence of other start- and end-of-sentence markers. E.g. UOrtmer . So wanne eyn claghe vor den rayt ku+omet [...] (furthermore, whenever a complaint before the council comes […]);.

iii. A binary feature which checks whether the token contains brackets, e.g., h(er)uorde (Herford).

For our experiments, we evaluated the performance of two different learning frameworks: memory-based learning and conditional random fields. Both learning methods can be described as classification-based supervised learning and have been shown to perform well on the task of part-of-speech tagging compared to decision-tree classifiers, Hidden Markov Models, etc. (e.g. Daelemans et al. 1999, Lafferty et al. 2001). The supervision lies in the fact that the learners are trained on the basis of annotated data. These data are manually annotated with POS information, which is taken from a predefined set of possible POS categories. During learning, both learners take as input training instances consisting of feature-value pairs (e.g. the standard and custom features as described above), followed by the annotated classification of that particular instance (as shown in Table 2). For the MBL classifier, the feature vector of each token includes a context window for every token, which consists of two tokens to the left and two tokens to the right. For CRF++, it is not necessary to include the local context in the feature vector, because CRF uses a feature template to instruct the learner to look at the labels of the tokens to the left and the right of the focus token. An example of such a feature vector is given in Table 3.

Table 3. Feature vector for the MBL classifier for the training sentence "MEn lest an der ol=den rethorica tu=liii."*

%	%	MEn	lest	an	0	0	1	0	0	0	0	3	me	en	men	men	men	DPIS
%	Men	lest	an	der	0	0	0	0	0	0	0	4	le	st	les	est	lest	VVFIN
MEn	lest	an	der	ol=den	0	0	0	0	0	0	0	2	an	an	an	an	an	APPR
lest	an	der	ol=den	rethorica	0	0	0	0	0	0	0	3	de	er	der	der	der	DDAR TA
an	der	ol=den	rethorica	tu=lii	0	0	0	0	0	1	0	6	ol	en	ol=	den	ol=den	ADJA
der	ol=den	rethorica	tu=lii	.	0	0	0	0	0	0	0	9	re	ca	ret	ica	rethorica	FM
ol=den	rethorica	tu=lii	.	$.$	0	0	0	0	0	1	0	6	tu	ii	tu=	lii	tu=lii	FM
rethorica	tu=lii	.	$.$	%	0	0	0	0	1	0	0	1	.	.	.	.	.	$;

tu=lii	.	$.$	%	%	1	0	0	0	1	0	0	3	$.	.$	$.$	$.$	$.$	!!ED!!

* The feature vector for the CRF classifier does not incorporate the first two columns nor the fourth and fifth column.

Memory-based learning (MBL) algorithms are called 'lazy learners' because they perform no generalization on the instance base they are trained on (Daelemans & van den Bosch 2005). All the instances are stored in memory, and new instances are classified by comparing them to the instance base, for example with a k-nearest neighbour algorithm. When a k-value of 1 is used, the classifier labels an unseen instance with its closest neighbour in the instance base. Various distance and feature weighting metrics can be used to determine which neighbour is closest. For larger values of k, some voting mechanism has to be applied to choose one class label from the nearest neighbours set. For our experiments, we used the MBL algorithms implemented in the *TIMBL* software package.

A Conditional Random Field (CRF) is a probabilistic classifier that is used to segment and label sequential data (such as a series of tokens), which makes it especially suitable for natural language processing tasks like part-of-speech tagging. Lafferty et al. (2001), for example, show that CRFs beat related classification models as well as HMMs on the POS tagging task. Similar results were recently obtained by Van de Kauter et al. (2013) for English, Dutch, French and German POS tagging and by Silfverberg et al. (2014) for English, Finnish, Czech, Estonian and Romanian.

CRFs take an input sequence X with its associated features, and try to infer hidden sequence Y, containing the class labels. For our experiments, *CRF++* version 0.57 was used. The classifier takes as input a template file that specifies the combinations of features it needs to consider. It also has a choice of several hyperparameters through which the behavior of the classifier can be tuned: a choice of regularization algorithm, balance between overfitting and underfitting, a cut-off threshold for feature frequency, and the number of threads (in case multi-threading is used).

4.3 Experimental data

We used three separate datasets for the evaluation: the texts Soester Schrae (henceforth Soest), Statuarrecht Rüthen (henceforth Rüthen) and Herforder Rechtsbuch (henceforth Herford). The three texts belong to the same genre (legal texts) and the same regional scribal variety (Westphalian (Altland)), but since language was not standardized during this period, the common scribal dialect only means that each city might have adopted features from the language of the regionally dominant city, Münster in this case. This is important for our purposes, because it means that a tagger trained on one city might still suffer from a performance drop when applied to another city. As the data sets are small, each has been split only into a train set and test set (80% for training and 20% testing), as presented in Table 4.

Table 4. Training and testing instances (tokens) for each dataset

	Training instances	**Testing instances**
Soest	6,593	1,649
Rüthen	5,427	1,377
Herford	12 968	3259

An analysis of the tag distribution of the different train sets (Figure 1) shows that even though Rüthen is the smallest set, it contains the largest number of categories, which makes it the most sparse one (i.e. there are multiple categories for which there are few instances for the classifier to generalize from). One of the main objectives for the tagger is that it needs to be robust to different kinds of input that may or may not conform to standard spelling.

PLEASEINSERT FIGURE 1 HERE

Figure 1. Tag distribution in the three training sets. The pink color refers to tags that do not occur in a given dataset. The darker the blue values, the higher the occurrence of a given tag in a given dataset

We already explained that the transcription scheme of the texts in the CHLG and the ReN projects often adds a lot of information to a token, e.g.:

su{R_n}der
vor#deghe=dinghen

ON_15
\FU_wesselere{R_n}\

This kind of transcription encodes the appearance of certain features of the manuscript, such as whether the lexical item is located in the main text, as super- or subscript, or in the margins. It may also indicate abbreviations in the text and to what lexical item they expand. Furthermore, it allows the original form of the manuscript to be recovered, which may provide important additional information, for instance where sentence boundaries are not clear – a common challenge in historical texts. However, the drawback of this transcription is that it is not very readable. This is why the POS annotation tool *CorA* offers two manners of representing the token, as the original transcription or as a simplified token without the transcription markers:

su{R_n}der —> sunder
vor#deghe=dinghen —> vordeghedinghen
ON_15 —> 15
\FU_wesselere{R_n}\ —> wesseleren

While in a normal resource-rich situation the simplified UTF8 spelling would offer less noisy input for the machine learner, we hypothesized that in a low-resource setting some cases of ambiguity might be resolvable with the additional information that the rich transcription offers.

We therefore prepared the datasets in three flavours with the purpose of establishing what would be the most informative input. The first dataset, called TRANS, consists of the token in its original transcribed form. The second dataset, called UTF8, takes the token in its UTF8 form as input, without additional annotation. The third group, TRANS_UTF8, presents the learner the transcription token and also the UTF8 form token as a feature.

5. Results and discussion

We conducted several experiments to assess the accuracy and robustness of the part-of-speech tagger for Middle-Low German. As a first step, we trained both MBL and CRF POS taggers for each city and tested them on data from the same city. In order to determine the cross-city robustness of the different city taggers, we gradually increased the training data and evaluated on each of the three city data sets. To have a clear view on which textual information was most helpful for tagging, we also performed feature selection using a wrapper-based genetic algorithm search methodology. Once having determined the optimal data set size (of course, taking into account the data we have at our disposal) and the optimal feature set, we conducted another experiment, in which we evaluated the performance of the tagger on another genre, viz. religious prose. Finally, we applied normalization in order to establish if the tagger can benefit from more regularized spelling.

5.1 In-domain experiments

In the first set of experiments, taggers were developed for each city using the two different learners *TiMBL* and *CRF++*, which were applied in their standard experimental parameter set-up for each of the three formats of the data (TRANS, UTF8 and TRANS_UTF8). The taggers were then tested on the 20% test data available for each dataset. We calculated two baselines (Table 5): a first baseline for which the most frequent tag from the training data ("NA" for all datasets) was taken as classification for all instances, and a second look-up baseline in which the tag for a given test token was looked up in the training data. For the latter baseline, we only report the best look-up baseline for the three varieties of the data sets.

Table 5. Baseline results

	Most frequent tag	**Look-up**
Soest	14.33	80.90
Rüthen	13.15	68.93
Herford	18.42	80.1

The results in Table 6 show that in comparison with *CRF++,* the *TiMBL* learner performs much worse on the Soest and Herford datasets (difference of around 5 percentage points) and slightly worse on the Rüthen dataset. Interestingly, the results on Soest and Herford are also up to 15% higher than those on Rüthen, which is likely caused by the higher sparsity of the latter dataset. Both systems show the best performance on the Soest TRANS_UTF8 set, and the Rüthen and Herford UTF8 sets.

Table 6. In-domain experimental results

	TiMBL		
	Transcription	**UTF-8**	**Transcription + UTF-8**
Soest	82.65	83.14	83.26
Rüthen	71.24	73.05	72.11
Herford	81.09	82.26	82.11
	CRF++		
Soest	87.62	87.38	87.68
Rüthen	73.56	74.58	73.85
Herford	85.54	86.37	86.25

5.2 Cross-city robustness

Given that POS taggers are trained on specific data sets, such as the above-mentioned Soest, Rüthen and Herford data sets, their models will typically also perform well on data with the same characteristics as the training data on which they based their models. In other words, a POS tagger trained on Soest would likely perform best on other texts from Soest. As it would require significant time and resources to produce manually POS annotated data for each single city, genre, etc., it is crucial that the developed tagger does not exhibit major performance drops when applied on unseen data.

In order to evaluate the out-of-domain robustness of the POS tagger, we conducted a set of cross-city experiments. The main objective was to find out what works best for our classifier: training on in-domain data or on a more diverse data set which incorporates a variety of material from different cities. In other words, is it better to have a small specifically tailored corpus, or does adding more data from other cities lead to a more robust and accurate tagger? We again trained POS taggers using the two learning frameworks and conducted 3 sets of experiments:

i. In a first set of experiments, we applied the already-trained single-city POS taggers on each of the other cities in order to compare the cross-city robustness of the taggers to their in-city performance. The robustness of the classifier was thus evaluated by exclusively training the classifier on out-of-domain data;

ii. In the second set of experiments, we added in-domain data to one of the out-of-domain datasets, thus increasing the number of training instances and also incorporating in-domain knowledge;

iii. In a third experiment, all available data were used for training.

Table 7. Cross-city robustness when training classifiers on out-of-domain data*

TiMBL									
	Soest			**Rüthen**			**Herford**		
	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8
Soest	82.65	83.14	83.26	64.34	65.14	65.35	66.70	67.96	67.41
Rüthen	74.71	75.01	73.19	71.24	73.05	72.11	68.42	69.83	69.07
Herford	72.95	73.49	74.77	67.53	68.91	68.33	81.09	82.26	82.11
CRF++									
	Soest			**Rüthen**			**Herford**		
	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8
Soest	87.62	87.38	83.68	63.76	64.48	63.25	71.80	72.84	72.59
Rüthen	80.04	80.41	79.65	73.56	74.58	73.85	74.13	75.85	75.29
Herford	78.65	79.01	78.65	70.00	71.38	70.37	85.54	86.37	86.25

* The grey results represent the in-domain experiments

Table 7 gives an overview of the cross-city robustness of the taggers that are exclusively trained on in-domain data. The columns indicate the text a tagger was trained on, the rows show the performance of that tagger on the test data, both from the same text (grey figures) and the two other texts (black figures). As expected, taggers perform worse on out-of-domain data, except for Rüthen, the smallest sub-corpus, for which the benefit of having more data outweighs that of having in-domain data.

In the second set of experiments (Table 8), we increased the training data with one out-of-domain set. When doing so, all three taggers increase in accuracy, the smallest gain being for the Rüthen dataset. Evidently, the results are lowest when no

training data from the city under consideration are added. The CRF classifier also consistently outperforms the MBL classifier.

Table 8. Cross-city robustness when training classifiers on both in-domain data and one out-of-domain data set*

TiMBL									
	Soest			**Rüthen**			**Herford**		
	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8
S+R	83.68	83.74	83.99	72.11	73.78	73.27	70.57	72.13	71.46
R+H	79.32	79.86	79.44	73.49	75.74	74.87	81.37	82.60	82.54
S+H	83.92	84.71	84.83	70.51	71.67	71.02	81.55	83.15	82.72
CRF++									
	Soest			**Rüthen**			**Herford**		
	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8	Trans	UTF8	Trans+ UTF8
S+R	**88.17**	88.29	88.11	75.23	75.67	75.52	77.10	78.12	77.47
R+H	84.11	85.56	84.41	**76.47**	**78.35**	**77.48**	**87.17**	**87.26**	**87.14**
S+H	87.93	**88.65**	**88.72**	73.20	75.45	73.85	86.19	86.86	86.89

* S = Soest, R = Rüthen, H = Herford

Adding all training sets (Figure 2) brings about the best results. When we compare the results in Figure 2 to the in-domain results in Table 6, we can observe a performance increase of 5% for all three datasets. Still, the results for Rüthen are the lowest, due to the combination of small size and a varied tagset.

PLEASE INSERT FIGURE 2 HERE

Figure 2. Performance on the three city datasets when all training data is used

5.3 Feature informativeness

As it is not immediately clear which features are more informative than others and what combinations work best, we performed feature selection using a wrapper-based approach exploiting genetic algorithms. There are different ways to select the optimal features for our POS task. One possible methodology consists in exhaustively testing all feature combinations, which is computationally very intensive. Another methodology

starts from one single feature and incrementally adds other features (forward selection) as long as this leads to performance increases or alternatively, starts from the full feature set and removes features one by one (backward elimination). A combination of these two directional search strategies is so-called bi-directional hillclimbing. A powerful alternative for feature selection which avoids searching the full feature space and at the same time is not bound by a certain search directionality, are genetic algorithms which start their search at different points in the search space (a so-called initial population) and seek the optimal feature combination through fitness-based selection (which individuals score best?) and operators such as mutation and cross-over to mutate and combine fit individuals into new generations.

We chose the latter approach to learn what features and learner hyperparameters lead to the optimal city tagger (Soest, Rüthen, Herford). Thus, for each learner (*CRF++* and *TiMBL*) we performed joint optimization on each city with each dataset type, resulting in nine experiments per learner. All optimization experiments were performed using the *Gallop* toolbox (Desmet & Hoste 2013). *Gallop* provides the functionality to wrap a complex optimization problem as a genome and to distribute the computational load of the GA run over multiple processors or to a computing cluster. It is specifically aimed at problems involving natural language. We set the initial population to 100 individuals set to run for 100 generations. The optimal features were based on the entire 80% dataset, validated with 5-fold cross-validation. The best fitness scores on the cross-validation data are reported in Table 9.

Table 9. Optimized in-domain experimental results (cross-validation)

	TiMBL		
	Transcription	UTF-8	Transcription + UTF-8
Soest	86.65	86.7	86.77
Rüthen	83.07	83.58	83.49
Herford	85.66	86.38	86.38
	CRF++		
Soest	87.64	**87.79**	87.76
Rüthen	84.52	**85.07**	84.85
Herford	86.69	**87.65**	87.57

Although these cross-validation results cannot be directly compared to the results reported in Table 6, we can observe that *CRF++* again achieves higher fitness values

than *TiMBL* for each city. Our main goal, however, was to determine which textual features improve tagging accuracy.

When analysing the results of the optimization, we followed Desmet (2014) and De Clercq (2015) and examined not only the one individual with the best fitness score, but also the individuals with the closest scores. For that procedure, we ranked the fitness scores and aggregated the best scores into five bins, based on the fitness score rounded to the fourth digit after the comma. Those individuals that had identical selected features and fitness were removed, so that only unique entries were analysed. In the resulting set of optimal feature sets, we noticed that on some occasions optimal fitness was reached when certain features were selected and sometimes when they were not. Given the small and varied datasets, we decided on three tiers of inclusion: features which were always selected, features which were selected more than 70 per cent of the time, and features which were selected between 50 per cent and 70 per cent of the cases.

Table 10. Selected features in the best individuals*

CRF++														
Soest														
trans	tok	parat	**vort**	**cap**	**dig**	**punc**	leng	pr2	s2	pr3	s3		*bra*	
utf8	tok	*parat*	*vort*	**cap**	dig	*punc*	leng	pr2	s2	pr3	s3	**hyp**	**bra**	
trans utf8		parat	**vort**	*cap*	dig	**punc**	**leng**	pr2	s2	pr3	s3	**hyp**		low
Rüthen														
trans	tok	parat	*vort*	**cap**	*dig*		**leng**	pr2		pr3	s3	hyp	*bra*	
utf8	tok	parat	*vort*	*cap*	dig			pr2	s2	pr3	s3	hyp	*bra*	
trans utf8		parat	*vort*	**cap**	dig	punc	leng		s2	pr3	s3	*hyp*		low
Herford														
trans	**tok**	*parat*	*vort*	cap	**dig**		leng	pr2	s2	pr3	s3	**hyp**	**bra**	
utf8	tok		**vort**	cap	dig	punc	leng	pr2	s2	pr3	s3	**hyp**	**bra**	
trans utf8		*parat*	*vort*		dig	punc	*leng*	pr2	s2	pr3	s3	*hyp*	*bra*	low

Features which are selected consistently (100%)
Features which are selected more than 70%
Features which are selected between 50% and 70%

* tok = token, para = paratext, vort = vortmer, cap = capitalization, dig = digit, punc = punctuation, leng = length, pr2 = bigram prefix, s2 = bigram suffix, pr3 = trigram prefix, s3 = trigram suffix, hyp = hyphen, bra = brackets, low = lowercase

Table 10 gives an overview of the selected features for the CRF learner that performed best on the three datasets. There are a number of conclusions we can draw: while the prefix, suffix, digit and length features seem to be consistently selected, this is less the case for the vortmer, capitalisation, punctuation, hyphen and bracket features.

5.4 Cross-genre robustness

We took the best-performing tagger and applied it on a fully unfamiliar dataset, the Spieghel der Leyen from Münster. As the dataset contains religious prose, albeit from the same regional scribal language, it differs sufficiently in style and vocabulary. The first 5,000 tokens of the text were manually tagged. Against these manual tags, we compared the automatic tags produced by three taggers:

i. the best non-optimized combined UTF8 tagger;
ii. an optimized version of the best tagger, where the features were selected through the genetic algorithm on the basis of the entire combined dataset of Soest, Rüthen and Herford, and;
iii. a version of the best tagger where the features were manually selected on the basis of the best features for each individual city tagger (Soest, Rüthen or Herford), as determined through the genetic algorithm. The UTF8 tagger was chosen, because of the three configurations (TRANS, UTF8 and TRANS_UTF8) UTF8 performs best on average across all experiments.

The non-optimized tagger tags with an accuracy of 75.93% but when optimized on the basis of the combined dataset, the performance on the new data drops nearly 1.5% to 74.63%. This indicates that the optimization on the basis of all datasets combined leads to overfitting. However, for the third tagger, we manually chose features from among the best-scoring taggers for each individual city. As each city benefitted from a different set of features, we went for the most salient ones (above 70%). Thus, the resulting tagger has more relaxed conditions and the accuracy in the out-of-domain dataset rises

to 77.11%. This is an improvement of 2.5% over the restrictive tagger and 1.2% over the non-optimized one. The features that were chosen manually are the token itself, capitalization, digit, hyphenation, brackets, word length and prefixes and suffixes with a length of 2 and 3 characters.

5.5 Error analysis

In order to have an overview of the remaining errors made by the best optimized tagger on the out-of-domain Spieghel der Leyen dataset, we manually checked whether there was any systematicity in the committed errors (Table 11). Overall, we observed two trends: problematic items are either tagged as one of five high-level tags (noun = NA, finite full verb = VVFIN, attributive adjective = ADJA, adverb = AVD, and preposition = APPR) or tagged as a semantically related category.

There are a number of labels that often occur instead of a small set of gold standard labels. Proper nouns, adjectives (ADJA, ADJS, ADJV), adverbs, finite verbs, infinitives, foreign material, and non-words are often falsely tagged as common nouns (NA). Proper and common nouns are commonly mistagged as infinitives.

Table 11. Frequent tagging errors

Gold Standard tag	**Incorrectly allocated tag**
NE ADJA ADJS ADJV AVD FM VVFIN VVINF	NA
NE VVFIN NA	VVINF
NA OA	VVFIN
PAVD	AVD
AVREL	AVD
PAVAP	APPR

KO*	KOUS
OA	!!ED!!

The most interesting cases are those in which the classifier makes a decision that makes sense semantically, but which affects the accuracy negatively because of the granularity of the tagset. For example, the tagset distinguishes between a tag for most types of adverbs (AVD) and a separate tag for relative adverbs (AVREL). The classifier often "erroneously" uses the more general AVD tag to also tag relative adverbs, which is counted as an error. Another example for the latter case is the group of pronominal adverbs. These adverbs take the form of discontinuous morphemes with one prepositional part and one pronominal part, and each part has a separate tag. For instance, in Example (1) below the clause:

(1) dar dit bok nicht af ne spricht
("about which (matter) this book does not speak",
lit. "there this book NEG about NEG speaks")

(Sachsenspiegel)

dar is the pronominal part and needs to tagged with PAVD, and *af* is the prepositional part and needs to be tagged with PAVAP. What happens in reality is that the prepositional part is tagged as a preposition (APPR) and the pronominal part PAVD - as an adverb AVD. This means that the classifier correctly disambiguates the function but does not account for the fact that those are two parts of a multi-word expression. This can be explained with the fact that PAVD occurs only ten times in the combined training data, but much more often in the Spieghel der Leyen religious prose dataset. Something similar happens when tagging conjunctions. The tagset distinguishes between coordinating conjunctions (KON), subordinating conjunctions (KOUS), and KO* which marks a conjunction (coordination or subordinating) which is a part of a multi-word expression. KO* does not appear at all in the training data, yet the tagger recognizes the respective tokens as conjunctions when it encounters them in the Spieghel der Leyen out-of-domain data. It tags those consistently as subordinating conjunctions KOUS.

6. Improving tagging accuracy: The impact of spelling normalization and morphological information

In order to explore what information could further improve the tagger, we set up two more experimental groups: one testing the contribution of spelling normalization to accuracy, and one testing the contribution of fine-grained morphological annotation. Part of the corpus has been labeled with fine-grained morphological tags, which give information about the inflection of each word (if applicable). Examples of inflectional categories in the tagset are: tense, mood, number, gender, case, etc. We use this morphologically annotated subset of the corpus to perform both of the aforementioned experimental tests (normalization and morphological analysis), so that we can compare directly how much each of them contributes to POS tagging accuracy.

6.1 Corpus subset and baseline

To create the training data for these experiments, two annotators independently tagged each of our core texts (Soest, Rüthen and Herford) and differences in the annotations were reconciled, resulting in three documents of the following sizes: Rüthen = 6,784 annotated tokens[2], Soest = 2,904 tokens, Herford = 5,367 tokens. We combined these three datasets into one, of which the training partition consist of 80% of each tagged text and the testing partition of the remaining 20%. This ensures that instances from each scribal dialect are present both in the training and testing partition.

The baseline experiment was to retrain the tagger on the reduced dataset using the best features and hyperparameters established in the experiments so far. The accuracy for this setup is 86.16%, a number that will serve as a reference point for all experiments henceforth. This experiment will be referred to as P1.

6.2 The effects of normalization on tagging accuracy

Although the MLG dialects were standardized up to a certain level, the MLG writing languages still display a lot of variation between the languages and even within one writing language. Spelling variation in particular impedes the creation of the POS tagger, since this variation causes sparseness of the input data on which the tagger has to train. This is illustrated in Example (2) from the Statuarrecht of Münster, in which we see two consecutive cases of the verb *sin* (to be) in the 3rd person plural (present tense), which are both spelled differently (*sint*/*synt*).

(2) […] al so vere als se da sint Synt se da nicht […]
("[…] as far as they are there. Are they not there, […]")
(Statuarrecht Münster)

To deal with these shortcomings, many NLP projects that focus on historical varieties begin with a normalization step, either because they aim to use resources built for contemporary varieties, or because they aim to train their own models on the historical variety and need to reduce sparseness. A notable example of such a spelling normalisation tool is *VARD* (Baron & Rayson, 2008), which has proven to improve performance on syntactic text processing (e.g. Schneider et al. 2015, Yang & Eisenstein 2016). *VARD* aids in bringing historical spelling closer to modern spelling which can lead to an improved performance of a contemporary POS tagger on historical text (Rayson et al. 2007). A similar experiment on our data with a state-of-the-art CRF tagger for German (Van de Kauter et al. 2013), however, revealed that all words were tagged as "Foreign Word", making this approach unfit for our purposes. Our focus is thus on normalization prior to training a dedicated POS tagger on historical text, in order to reduce sparseness on the input tokens.

There are two main approaches to normalization: rule-based approaches and data-driven approaches. The rule-based systems i.a. rely on mapping schemes from historical to modern spelling (e.g. *VARD*) or apply a Levenshtein similarity approach (Pettersson et al. 2013). While these rule-based approaches make the best use of expert knowledge, they typically need much work in order to become robust and developing them is time-intensive. Data-driven approaches, on the other hand, derive their knowledge from large amounts of normalized data. A recent popular data-driven

methodology for normalization is the use of character-based statistical machine translation (Schulz et al. 2016, Pettersson et al. 2014), which partly overcomes the requirement for large amounts of data as the system works at the character level.

Our main aim was to make the tagger as robust to spelling variation as possible and as independent from external resources as possible. That was achieved through the use of our special features such as brackets, editorial annotations, capitalization and character n-grams. After achieving the highest possible scores with that setup and feature experimentation, we constructed a normalization script that would help us assess how much more advantage normalization might give us. Given our rather small dataset, we opted for a rule-based normalization methodology.

For the normalization script, written in *Python*, we defined a set of 26 rules and exceptions affecting about 60 spelling variants. Complementary variants were all listed with their number of occurrences and their relative frequencies. We decided to change each spelling variant to the most common complementary variant. In a next step, all the words containing a variant for which the rule applied were extracted to make sure the rule performed well. In that way, some exceptions were discovered and added to the script. An example of a rule with its exceptions is the orthographic variation in representing */e:/*, for which the most common spelling is <*ey*>. The other possible spelling options, <*ei*> and <*ee*> are therefore changed to this spelling. There is however an exception that has to be added: if <*ei*> or <*ee*> are following a */b/* or */g(h)/* at the beginning of a word, the rule should not be applied, since in that case we are likely dealing with a syllable boundary after the prefix of the past participle, so between *e* and *e* or *e* and *i* (e.g. in *beendet*, “finished”). The script was tailored to the three texts used in the experiments, and all strings complying with a rule or exception rule in the script are affected. Normalization accuracy was not evaluated on other texts, since the main aim was to preprocess the three texts under study and measure the maximal improvement in POS tagging performance, given gold-standard normalization.

When applying the normalization script to each city dataset in UTF8 form in its entirety, we note that the script affects 8.09% of the tokens in Soest, 7.70% of the tokens in Herford, and 13.22% of the tokens in Rüthen.

Table 12. Effect of normalization on the three complete datasets, as expressed by token count and as percentage

Complete dataset	Dataset size (tokens)	Normalized tokens (token count)	Normalized tokens (percentage)
Soest	8,241	667	8.09%
Rüthen	6,784	897	13.22%
Herford	16,228	1,250	7.70%

For the experiments, normalization was applied on the reduced dataset as described in Section 6.1. Normalization affected 8.77% of the tokens in the training dataset, and 8.88% of the tokens in testing.

We evaluated two ways of applying the normalization: (i) by using the normalized token as a feature in addition to the original token, and (ii) instead of the original token. Moreover, we again used a setup with strict (probably overfitting) features and a setup with relaxed feature conditions, where the most suitable features were manually selected from each optimized city tagger. The four sets of experiments confirm that the relaxed set of features is at least as robust as the strict feature set. The results of the strict set experiments and the first relaxed set experiment are comparable to those of the non-normalized best tagger. Only the combination of a normalized input token and relaxed features brings about 1% improvement.

Table 13. POS accuracy with different normalization setups

	Strict feature set		**Relaxed Feature set**	
	Norm. token as extra feature	Norm. token instead	Norm. token as extra feature	Norm. token instead
Reduced dataset	86.26	86.95	86.16	87.12

An error analysis on the best normalized optimized tagger against the best non-optimized tagger showed that for larger and less variable test sets (Soest and Herford), none of the normalized lexical items are tagged differently (i.e. if the non-normalized tagger tags correctly, so can the normalized one, and if the non-normalized tagger tags the item wrongly, so does the normalized). From this, we can draw the conclusion that given a sufficiently large dataset with less variable tags the tagger can cope with spelling variation. The errors that remain are the result of linguistic phenomena or inconsistencies in the tagset.

6.3 The effects of morphological information on POS tagging

Another possible way of improving tagging accuracy is by including fine-grained morphological information. There are different ways in which it can be obtained, and different ways in which it can be incorporated. With these experiments, we address two questions: (i) can morphological information improve POS tagging, and (ii) can POS information be used to generate accurate morphological annotation that can be used for further experiments?

It is important to know that trying to generate POS and morphological information in one go from the data, like *TreeTagger* and other taggers using the STTS tagset do, leads to serious sparseness issues. While the POS-only tagset is under 100 tags, a combined POS-and-morphology tagset grows to a size of nearly 500. This means that in a training document of less than 13,000 tokens, many of the tags would only be seen once. Indeed, when we experimented with predicting full tags from the token and features, the accuracy was only 70.4% on the test set.

The first step in predicting POS is establishing an upper bound and estimating how much adding gold standard morphological information can help performance. We tested two setups in which the POS tag is predicted using the token, optimized POS features, and gold standard morphology. In setup C1, the morphology is incorporated into the feature vector by decomposing it into components (e.g. Neut, Fem, Dat, Gen, Sg, etc.) and coding the presence of each possible component with a binary value. In setup C2, on the other hand, the morphological information is coded as a string (e.g. Neut.Dat.Sg). When tested on the test set, the two experiments set the upper bound at 91.40% and 90.65% for C1 and C2, respectively. This indicates that decomposing the morphological information into binary features is worthwhile. Predicting the POS from only the token and the gold standard morphology results in an accuracy of 87%, or no improvement over predicting POS from token and features only.

The second step is predicting POS using predicted morphological information, in order to estimate how much deterioration in the reliability of the morphological information is acceptable before it impacts POS tagging accuracy.

We predict morphology from two initial setups that we provisionally call M1 and M2. In M1 we predict the morphological tag from the token, POS features and the gold standard POS tag with an accuracy on the test set of 79.1%. In M2 we predict it only from the token and POS features, achieving an accuracy of 75%. Using predicted morphological information from M1 (79% accurate prediction) together with only the token leads to a POS tagging accuracy of 86.2%. When using morphological information from M2 (75% accuracy), the POS tagging accuracy already drops to 80%, which is under the baseline that we established with P1.

In a final set of experiments, we compare POS prediction when the input is token, POS features and predicted morphology from M1 and M2. The results are 90.89% and 86.16%, respectively. Clearly, the first experiment, based on 79% correct morphology, nearly reaches our upper bound.

While these morphological experiments constitute a proof of concept at this stage, the fact that imperfect predicted morphology is showing to have positive impact on POS accuracy opens up the path for more targeted work in this direction.

7. Conclusions

We reported on a data-driven approach to build a robust Part-of-Speech tagger for Middle Low German. We obtained an in-domain accuracy of up to 87.7% when evaluating on three datasets from different cities, all three belonging to the same genre (legal texts) and the Westphalian variety, but nevertheless showing considerable spelling variation. While a drop of performance was observed in the cross-city experiments, the best tagging results were obtained by adding both in-city and cross-city training data. In order to assess feature informativeness, we performed feature selection using a wrapper-based approach exploiting genetic algorithms, leading to the selection of a set of features such as bigram and trigram prefix and suffix features, which was consistently selected in the different experiments. As the tagger suffered from overfitting on the in-domain data, a more relaxed feature and hyperparameter setup was used for the remainder of the experiments. In a cross-genre robustness experiment on a

religious prose dataset, we obtained a 77.1% accuracy and observed that part of the errors made sense semantically and were mainly caused by the granularity of the tagset.

We also investigated the role of normalization and of fine-grained inflectional morphological analysis in potentially improving the tagger's accuracy. To measure the impact of spelling normalization on tagging accuracy, a rule-based normalization script was written and two sets of experiments were conducted, viz. one in which the normalized token was added to the original token and one in which it replaced the original token. Interestingly and contrary to previous findings, we could observe that the tagger was quite robust to spelling variation. Finally, in a set of experiments in which we measured the impact of both perfect and predicted morphological information on POS tagging accuracy, we could observe that even imperfect predicted morphological information has a positive effect on tagging performance.

Comparing our approach to earlier results reported for POS tagging MLG texts with *RFtagger* (Barteld et al. 2015), which only reached an accuracy of maximally 75.7% ± 1.6, we can conclude that our approach compares favorably to the results reported earlier (but on different texts) both with respect to accuracy, flexibility and robustness.[3]

Notes

1. While it is possible to encode in the template that the CRF algorithm also looks at the tokens themselves (i.e. using true ngrams as opposed to label-ngrams), this is not the default CRF behavior and some exploratory experiments showed that it affects the accuracy negatively.

2. The full text of Statuarrecht Rüthen. The number of tokens differs from the one reported for the POS experiments in the previous sections, because some retokenization occurred during the morphological annotation, particularly where clitics were concerned.

3. Barteld et al. (2007) both train and test RFtagger within one domain concerning scribal language and period, namely two North Low Saxon prose texts (one religious, one literary) from around 1500 (1480 and 1502). Without regularization, the maximum accuracy of RFtagger

trained on the other text applied to the other is 73.5% ± 1.9, with regularization, it is 75.7% ± 1.6.

References

Baron, A., & Rayson, P. (2008, August). *VARD2: A tool for dealing with spelling variation in historical corpora*. Paper presented at Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, UK.

Barteld, F., Schröder, I., & Zinsmeister, H. (2015). Unsupervised regularisation of historical texts for POS tagging. In F. Mambrini, M. Passarotti & C. Sporleder (Eds.), *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)* (pp. 3-12). Polish Academy of Sciences: Institute of Computer Science.

Bennett, P., Durrell, M., Scheible, S., & Whitt, R. J. (2010). Annotating a historical corpus of German: A case study. In *Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards* (pp. 64-68). European Language Resources Association.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biebersteadt, A. (2015). Variablenlinguistische Beobachtungen zu den mittelniederdeutschen Schreibsprachen des südlichen Ostseeraumes: Wismar und Stralsund als Beispiele. In: H. U. Schmid & A. Ziegler (Eds.) *2015: Jahrbuch für Germanistische Sprachgeschichte. Bd. 6: Deutsch im Norden* (pp. 88-115). Berlin/New York: De Gruyter.

Bollmann, M., Petran, F., Dipper, S., & Krasselt, J. (2014). CorA: A web-based annotation tool for historical and other non-standard language data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (pp. 86-90).

Braunmüller, K. (1996). Forms of language contact in the area of the Hanseatic League: Dialect contact phenomena and semicommunication. *Nordic Journal of Linguistics, 19*(2), 141-154.

Braunmüller, K. (2002). Language contact during the Old Nordic period I: With the British Isles, Frisia and the Hanseatic League. In O. Bandle, K. Braunmüller, E. H. Jahr, A. Karker, H.-P. Naumann & U. Teleman (Eds.), *The Nordic Languages: An*

International Handbook of the History of the Nordic Germanic Languages, Volume 1 (pp. 1028-1039). Berlin/New York: De Gruyter.

Breitbarth, A., Walkden, G., & Watts, S. (2011 April). *A Corpus for Middle Low German*. Paper presented at New Methods in Historical Corpora, Manchester, UK.

Breitbarth, A., Walkden, G., & Watts, S. (2012 April). *Building a corpus for Middle Low German: Notes and queries*. Paper presented at the Forum for Germanic Language Studies (FGLS10), Sheffield, UK..

Daelemans, W., Van den Bosch, A., & Zavrel, J. (1999). Forgetting examples is harmful in language learning. *Machine Learning, 34*(1-3), 11-43.

De Clercq, O. (2015). *Tipping the scales: exploring the added value of deep semantic processing on readability prediction and sentiment analysis* (Unpublished doctoral dissertation). Ghent University, Ghent, Belgium.

Desmet, B., Hoste, V., Verstraeten, D., & Verhasselt, J. (2013). *Gallop Documentation*, (LT3 Technical Report - LT3 13.03).

Desmet, B. (2014). *Finding the online cry for help: Automatic text classification for suicide prevention* (Unpublished doctoral dissertation). Ghent University, Ghent, Belgium.

Diel, M., Fisseni, B., Lenders, W., & Schmitz, H.-C. (2002). XML-Kodierung des Bonner Frühneuhochdeutschkorpus. Bonn: IKP-Arbeitsbericht NF 02.

Dipper, S. (2015). Annotierte Korpora für die Historische Syntaxforschung: Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch. *Zeitschrift für Germanistische Linguistik, 43*(3), 516-563.

Dipper, S., Donhauser, K., Klein, T., Linde, S., Müller, S., & Wegera, K. P. (2013). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics, 28*(1), 85-137.

Fisseni, B., Schmitz, H.-C., & Schröder, B. (2007). FnhdC/HTML und FnhdC/S. *Sprache und Datenverarbeitung, 1-2/2007*, 67-69.

Geyken, A., Haaf, S., Jurish, B., Schulz, M., Steinmann, J., Thomas, C., & Wiegand, F. (2011). Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, *20./21, September 2010, Beiträge der Tagung, 2., ergänzte Fassung* (pp. 157-161).

Kroch, A., Taylor, A., & Ringe, D. (2000). The Middle English verb-second constraint: A case study in language contact and language change. In S. Herring, P. van Reenen & L. Schøsler (Eds.) *Textual Parameters in Older Languages* (pp. 353-392). Amsterdam/Philadelphia: Benjamins.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 282-289). San Francisco, CA: Morgan Kaufmann.

Linde, S., & Mittmann, R. (2013). Old German reference corpus: Digitizing the knowledge of the 19th century. In P. Bennett, M. Durrell, S. Scheible, R. J. Whitt (Eds.) *New Methods in Historical Corpora* (pp. 235-246). Tübingen: Narr Verlag.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT press.

Marcus, M. P., Santorini B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313-330.

Martineau, F. (2005). Modéliser le changement: Les voies du français/Modelling change: The paths of French. Ottawa: University of Ottawa. Retrieved from www.voies.uottawa.ca/corpus_pg_en.html (last accessed March 2017).

Moon, T., & Baldridge, J. (2007). Part-of-speech tagging for Middle English through alignment and projection of parallel diachronic texts. In *Proceedings of EMNLP/CONLL-2007* (pp. 390-399).

Peters, R. (1973). Mittelniederdeutsche Sprache. In J. Goossens (Ed.) *Niederdeutsch – Sprache und Literatur. Bd. 1: Sprache* (pp. 66-115). Neumünster: Wachholtz.

Peters, R. (2003). Variation und Ausgleich in den mittelniederdeutschen Schreibsprachen. In M. Goyens & W. Verbeke (Eds.), *The Dawn of the Written Vernacular in Western Europe* (pp. 427-440). Leuven: Leuven University Press.

Peters, R., & Fischer, C. (2007). Der 'Atlas spätmittelalterlicher Schreibsprachen des niederdeutschen Altlandes und angrenzender Gebiete'. In L. Czajkowski, C. Hoffmann, H.U. Schmid (Eds.) *Ostmitteldeutsche Schreibsprachen im Spätmittelalter* (pp. 23-33). Berlin: De Gruyter.

Pettersson, E., Megyesi, B., & Nivre, J. (2013). Normalisation of historical text using context-sensitive weighted Levenhstein distance and compound splitting. In *Proceedings of the 19th Nordic Conference on Computational Linguistics* (NoDaLiDa 2013) (pp. 163-179). Linköping: Linköping Electronic Conference Proceedings 85.

Pettersson, E., Megyesi, B., & Nivre, J. (2014). A multilingual evaluation of three spelling normalization methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH 2014)* (pp. 32-41). Gothenburg: Association for Computational Linguistics.

Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the bard: Evaluating the accuracy of a modern POS tagger on early modern English corpora. In *Proceedings of Corpus Linguistics 2007*. Birmingham: University of Birmingham, UK.

Rögnvaldsson, E., & Helgadóttir, S. (2011). Morphosyntactic tagging of Old Icelandic texts and its use in studying syntactic variation and change. In C. Sporleder, A. van den Bosch, K. Zervanou (Eds.) *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series* (pp. 63-76). Berlin: Springer.

Sanders, W. (1982). Sprachgeschichtliche Grundzüge des Niederdeutschen. Vandenhoeck + Ruprecht Gm.

Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011a). A gold standard corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V 2011)* (pp. 124-128). Association for Computational Linguistics.

Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011b). Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, pp. 19-23. Portland, OR: Association for Computational Linguistics.

Schiller, A., Teufel, S., & Thielen, C. (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universities of Stuttgart and Tübingen, 66. Retrieved from www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (last accessed March 2017).

Schmid, H., & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008) - Volume 1* (pp. 777-784). Manchester: Association for Computational Linguistics.

Schneider, G., Lehman, H. M., & Schneider, P. (2015). Parsing early and late modern English corpora. *Literary and Linguistic Computing, 30*(3), 423-439.

Schröder, I. (2014). Neue Perspektiven für die mittelniederdeutsche Grammatikographie. *Jahrbuch für germanistische Sprachgeschichte, 5*(1), 150-164.

Schulz, S., De Pauw, G. De Clercq, O., Desmet, B., Hoste, V., Daelemans, W., & Macken, L. (2016). Multimodular Text Normalization of Dutch User-Generated Content. *ACM Transactions on Intelligent Systems and Technology (TIST), 7*(4), 1-22.

Taylor, A., Warner, A. Pintzuk, S., & Beths, F. (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). University of York: Department of Language and Linguistic Science. Retrieved from

www.ling.upenn.edu/mideng/ppcme2dir/YCOE/YcoeHome.htm (last accessed March 2017).

Silfverberg, M., Ruokolainen, B., Lindén, K., & Kurimo, M. (2014). Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 259-264). Baltimore, MD.

Sukhareva, M., & Chiarcos, C. (2016). Combining ontologies and neural networks for analyzing historical language varieties: A case study in Middle Low German. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & Stelios Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris: European Language Resources Association (ELRA). Retrieved from www.lrec-conf.org/proceedings/lrec2016/summaries/822.html (last accessed March 2017).

Tophinke, D. (2009). Vom Vorlesetext zum Lesetext: Zur Syntax mittelniederdeutscher Rechtsverordnungen im Spätmittelalter. In A. Linke, & H. Feilke (Eds.), *Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt* (pp. 161-186). Tübingen: Niemeyer.

Tophinke, D. (2012). Syntaktischer Ausbau im Mittelniederdeutschen. Theoretisch-methodische Überlegungen und kursorische Analysen. *Niederdeutsches Wort, 52*, 19-46.

Tophinke, D., & Wallmeier, N. (2011). Textverdichtungsprozesse im Spämittelalter: Syntaktischer Wandel in mittelniederdeutschen Rechtstexten des 13.–16. Jahrhunderts. In S. Elspaß & M. Negele (Eds.) *Sprachvariation und Sprachwandel in der Stadt der Frühen Neuzeit* (pp. 97-116). Heidelberg: Winter.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal, 3*, 103-120.

Walkden, G. (2016). The HeliPaD: A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics, 21*(4), 559-571.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). Icelandic parsed historical corpus (IcePaHC) (Version 0.9). Available at www.linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_%28IcePaHC%29 (last accessed March 2017).

Yang, Y., & Eisenstein, J. (2016). Part-of-speech tagging for historical English. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, San Diego.

Authors' addresses

Mariya Koleva
LT3 Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45
9000 Ghent
Belgium

m.s.koleva@outlook.com

Melissa Farasyn
Department of Linguistics (ΔiaLing)
Ghent University
Blandijnberg 2
9000 Ghent
Belgium

melissa.farasyn@ugent.be

Bart Desmet
LT3 Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45
9000 Ghent
Belgium

bart.desmet@ugent.be

Anne Breitbarth
Department of Linguistics (ΔiaLing)
Ghent University
Blandijnberg 2
9000 Ghent
Belgium

anne.breitbarth@ugent.be

Veronique Hoste
LT3 Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45
9000 Ghent
Belgium

veronique.hoste@ugent.be