Rule-Based Normalisation of Historical Text – a Diachronic Study

Eva Pettersson[†], **Beáta Megyesi and Joakim Nivre**
Department of Linguistics and Philology
Uppsala University
[†]Swedish National Graduate School
of Language Technology
firstname.lastname@lingfil.uu.se

Abstract

Language technology tools can be very useful for making information concealed in historical documents more easily accessible to historians, linguists and other researchers in humanities. For many languages, there is however a lack of linguistically annotated historical data that could be used for training NLP tools adapted to historical text. One way of avoiding the data sparseness problem in this context is to normalise the input text to a more modern spelling, before applying NLP tools trained on contemporary corpora. In this paper, we explore the impact of a set of hand-crafted normalisation rules on Swedish texts ranging from 1527 to 1812. Normalisation accuracy as well as tagging and parsing performance are evaluated. We show that, even though the rules were generated on the basis of one 17th century text sample, the rules are applicable to all texts, regardless of time period and text genre. This clearly indicates that spelling correction is a useful strategy for applying contemporary NLP tools to historical text.

1 Introduction

Digitalised historical text is a rich source of information for researchers in humanities. However, there is a lack of language technology tools adapted to old texts, that could make the information more easily accessible. As an example, the work presented in this paper has been carried out in close cooperation with historians in the *Gender and Work* project (Ågren et al., 2011). This project aims at exploring how men and women supported themselves in the Early Modern Swedish period (1550–1800). Information on this is manually searched for in documents from this period, and stored in a database. The project has shown that work activities are most often described as a verb and its complements. Automatically extracting all the verbs in a text would be a rather trivial task for contemporary text, using standard NLP tokenisation and tagging. However, for historians and other researchers in humanities, the manual way of searching for information is still often the only alternative.

Developing NLP tools for historical text is a great challenge, since old texts vary greatly in both spelling and grammar between different authors, genres and time periods, and even within the same text, due to the lack of spelling conventions. The texts that we have studied are generally written in a spoken language fashion, with less distinct boundaries between sentences, and a spelling that to a larger extent reflects the phonetic form of the word. Sentences with 50–100 words or more are not unusual. The Swedish language was also strongly influenced by other languages at this time. Evidence of this is the placement of the finite verb at the end of subordinate clauses in a German-like style not usually found in modern Swedish texts.

Another problem in developing language technology tools specifically aimed at handling historical text, is the shortage of linguistically annotated historical data. One way of getting around this problem is to use existing NLP tools trained for handling contemporary text, and normalise the historical input text to a more modern spelling be-

Proceedings of KONVENS 2012 (LThist 2012 workshop), Vienna, September 21, 2012

fore applying the tools. Pettersson et al. (2012) showed that a relatively small set of hand-crafted normalisation rules had a large positive impact on the results of tagging and parsing of historical text using contemporary NLP tools. We are interested in how well this method works for texts from different time periods and genres. In this paper, we present a case study, where we evaluate the impact of these normalisation rules on 15 different Swedish texts ranging from 1527 to 1812, within the genres of court records versus church documents. Our hypothesis is that the normalisation rules will work the best for texts from the same time period as the text sample used for developing the normalisation rules, and within the same text genre, i.e. court records from the 17th century. However, we hope that the rules will be of use also for both older and younger texts, as well as for other text types. Furthermore, we assume that error reduction as regards normalisation will be larger for older texts, since these differ from contemporary language to a larger extent, whereas the overall normalisation accuracy probably will be higher for younger texts.

From the evaluation we see that the rules are indeed applicable to all texts in the corpus, substantially increasing the proportion of normalised words in the text, as well as improving tagging and parsing results. Part of the explanation for this is that the same types of spelling changes occur in texts from all time periods studied, only with a slightly different frequency distribution.

2 Related Work

Sánchez-Marco et al. (2011) tried a strategy for adapting an existing NLP tool to deal with Old Spanish. Adaptation was performed on the basis of a 20 million token corpus of texts from the 12th to the 16th century, and included expansion of the dictionary, modification of tokenisation and affixation rules, and retraining of the tagger. The dictionary expansion had the highest impact on the results, and was performed by automatically generating word forms through mapping old spelling variants to their contemporary counterparts. The tagger was then retrained based on a gold standard of 30,000 tokens, where the tokens were first pre-annotated with the contemporary tagger, and then manually corrected. The evaluation of the final tagger showed an accuracy of 95% in finding the right part of speech, and 90% accuracy in finding the complete morphological tag.

Bollmann et al. (2011) tried a rule-based approach to normalisation of texts from Early New High German (14th to 16th century). In this approach, rules were automatically derived from word-aligned parallel corpora, i.e. the 1545 edition of the Martin Luther Bible aligned with a modern version of the same Bible. Since Bible text is rather conservative in spelling and terminology, approximately 65% of the words in the old Bible version already had an identical spelling to the one occurring in the modern version. To cope with non-identical word forms, normalisation rules were generated from the word alignment results, by means of Levenshtein edit distance, and the rules were sorted by frequency of occurrence. In order not to generate non-words, only word forms that could be found in the modern version of the Bible were accepted. Other word forms produced by the rules were discarded, leaving the old spelling preserved. This method proved successful, increasing the proportion of words with a modern spelling from 65% to 91%.

Oravecz et al. (2010) included a standardisation/normalisation step in their work on semi-automatically annotating a corpus of Old Hungarian. Normalisation was performed using a noisy channel model combined with morphological analysis filtering and decision tree reranking. Combining these methods, they reached a normalisation precision of 73%.

Pettersson et al. (2012) used a relatively small set of hand-crafted normalisation rules for improving tagging and parsing of historical text. The aim of this study was to automatically extract verbs and their complements from Early Modern Swedish texts (1550–1800). A part-of-speech tagger trained on contemporary Swedish text was used for verb identification, whereas verbal complements were extracted based on output from a dependency parser. Spelling variation in the input text was handled by a set of 29 hand-crafted rules, produced on the basis of a sample text from the 17th century. The method was evaluated on a corpus of 400 sentences extracted from the period 1527–1737. For verb extraction based on tagging, recall increased from 60% to 76% when normal-

isation rules were applied to the input text before running the tagger. Likewise, the proportion of fully or partially extracted verbal complements based on parsing increased from 53% to 56%.

3 Normalisation Method

In our experiments, we will use the setup described in Pettersson et al. (2012), where a set of 29 hand-crafted normalisation rules are used for transforming the spelling into a more modern version. The normalisation rules were developed on the basis of two sources:

1. Spelling changes documented in the context of the reformed Swedish spelling introduced in 1906 (Bergman, 1995).

2. A text sample from *Per Larssons dombok*, a selection of court records from 1638 (Edling, 1937).

Since 1906, no larger spelling changes have been employed in the Swedish language. With this reform, the spelling for the *t* sound was simplified from *dt* to a single *t*, as in *varidt→varit* (”been”). Likewise for the *v* sound, the superfluous letters *h* or *f* were dropped, as in *hvar→var* (”was”) and *skrifva→skriva* (”write”). Also, the use of an *f* for denoting a *v* sound was abandoned, turning for example *af* into *av* (”of/off”).

As for the empirical rules based on the 17th century text sample, these include:

- substitution of letters to a phonologically similar variant, such as q→k in *qvarn→kvarn* (”mill”) and z→s in *slogz→slogs* (”were fighting”)

- deletion of repeated vowels, as in *saak→sak* (”thing”)

- deletion of mute letters, such as j in *vijka→vika* (”fold”), and b in *dömbdes→dömdes* (”was sentenced”)

- normalisation of spelling influenced by other languages, mainly German, as in *schall→skall* (”shall”)

The current normalisation scheme has no mechanism for handling word forms that are no longer in use. These will be treated like any other word by the normalisation algorithm.

4 Hypothesis

Our initial hypothesis is that the set of normalisation rules described in section 3, will work better for 17th century court records than for texts from other centuries and within other genres, since most of the rules have been developed on the basis of a text sample from 1638. Since there were no standardised spelling conventions during this period, we also believe that there will be differences in the applicability of the rules between texts from different authors and within different text genres.

A closer look at the gold standard however shows that the most frequent types of spelling changes between the old and the modern text are more or less the same, regardless of time period and text type. Table 1 shows the 10 most frequent changes (at a character level) between the raw historical text and the manually normalised gold standard. In this table, a minus sign means that a letter needs to be removed to transform the spelling into the modern spelling given in the gold standard. For example *-h* is performed for transforming *åhr* into *år* (”year”). In the same way, a plus sign denotes that a letter is inserted to create the modern spelling. For example, *+dbl* means that a letter is duplicated, as when transforming the spelling *alt* into the contemporary spelling *allt* (”everything”). The ”/” sign means that a letter needs to be replaced by another letter, most commonly a phonologically close letter. This can be illustrated by the change *w/v*, transforming for example *beswär* into *besvär* (”trouble”).

As seen in the table, the most frequently occurring changes are present in texts from all centuries represented in the corpus, and in both court records and church documents. The frequency distribution differs only slightly, so that for example the deletion of *-h* is the most common transformation seen in texts from the 16th and 17th century, whereas this deletion is only the second most frequent in 18th century text. This information makes us believe that the normalisation rules may have a significant impact on all texts, not only 17th century court records as stated in our initial hypothesis.

16 cent	17 cent	18 cent	Court	Church
-h	-h	w/v	-h	-h
w/v	-dbl	-h	-dbl	w/v
e/ä	w/v	+dbl	w/v	-dbl
-f	-f	-e	-e	e/a
-dbl	-e	-f	-f	-f
e/a	e/a	e/ä	+dbl	e/ä
-e	+dbl	f/v	e/ä	-e
+dbl	e/ä	e/a	e/a	+dbl
u/v	f/v	-dbl	-i	c/k
c/k	i/j	c/k	f/v	i/j

Table 1: Top 10 changes at a character level, for transforming raw historical text into the manually normalised gold standard spelling. 16 cent = 16th century texts (1527–1593). 17 cent = 17th century texts (1602–1689). 18 cent = 18th century text (1728–1812). Court = Court records. Church = Church documents. Dbl = Double letter.

Court Records			
Name	**Year**	**Total**	**Sample**
Östra Härad	1602–1605	38,477	2,069
Vendel	1615–1645	64,977	2,509
Per Larsson	1638	12,864	2,987
Hammerdal	1649–1686	75,143	1,859
Revsund	1649–1689	113,395	2,328
Stora Malm	1728–1741	458,548	1,895
Vendel	1736–1737	61,664	3,450
Stora Malm	1742–1760	74,487	2,336
Stora Malm	1761–1783	66,236	1,825
Stora Malm	1784–1795	58,738	1,378
Stora Malm	1796–1812	47,671	1,683
Church Documents			
Name	**Year**	**Total**	**Sample**
Västerås	1527	14,149	3,709
Kyrkoordning	1571	60,354	2,246
Uppsala Möte	1593	34,877	1,184
Kyrkolag	1686	35,201	2,086
Total	**1527–1812**	**787,122**	**33,544**

Table 2: Corpus distribution, given in number of tokens in the documents. Total = Number of tokens in the whole corpus. Sample = Number of tokens in the gold standard sample.

5 Experimental Setup

The normalisation rules described in section 3, and their impact on tagging and parsing results, are to be evaluated taking different text types and time periods into account. The corpus used for evaluation consists of 15 texts, ranging from 1527 to 1812. The text types covered are court records and church documents. 10 out of these 15 texts are the same as the ones used for evaluation by Pettersson et al. (2012). In total, there are 787,122 tokens in the corpus. From this corpus, a gold standard has been extracted, comprising 40 randomly selected sentences from each text, i.e. in total 600 sentences. The proportion of tokens in each text as a whole, and in the gold standard part of the text, is illustrated in table 2.[1]

For each text in the gold standard corpus, evaluation includes 1) the proportion of correctly normalised tokens, and the error reduction achieved through normalisation, 2) the accuracy of verb identification (tagging), before and after normalisation, and 3) the accuracy of verbal complement extraction (parsing), before and after normalisation. Error reduction has been calculated by the following formula:

$$\frac{\text{CorrectAfterNormalisation} - \text{CorrectBeforeNormalisation}}{\text{IncorrectBeforeNormalisation}}$$

where *CorrectAfterNormalisation* is the percentage of tokens with an identical spelling to the modern version *after* the normalisation rules have been applied, *CorrectBeforeNormalisation* is the percentage of tokens with an identical spelling to the modern version *before* the normalisation rules have been applied, and *IncorrectBeforeNormalisation* is the percentage of tokens differing in spelling from the modern version before the normalisation rules have been applied.

To be able to evaluate the proportion of correctly normalised tokens, each token in the gold standard part of the corpus was manually assigned its modern spelling equivalent. For tagging and parsing, there is currently no gold standard available for the texts used in the gold standard corpus. However, since the overall aim of the research project (*Gender and Work*, described in section 1) is to extract verbs and their complements from historical text, all the verbs and their comple-

[1] The number of tokens differs slightly from those presented in Pettersson et al. (2012), since the current numbers are calculated on tokenised text, whereas the previous numbers were calculated on the "raw" source text.

ments in the gold standard have been manually annotated. Therefore we may indirectly evaluate tagging accuracy by comparing the verb tags produced by the tagger to the words marked as verbs in the gold standard. Likewise, parsing accuracy may be evaluated by comparing the verbal complements assigned by the parser to the verbal complements given in the gold standard.

6 Results and Discussion

6.1 Normalisation

Table 3 presents the proportion of tokens in the historical texts with the same spelling as in the manually modernised gold standard, before and after normalisation.[2] This table also shows the error reduction for each test text, i.e. the percentage of correctly normalised tokens that were not originally identical to the modern spelling (see further section 5 for a definition of the calculation of error reduction).

Since the normalisation rules are based on a text sample from 17th century court records, it could be expected that the rules would not be applicable for other time periods and text types. However, the results presented in table 3 show that this set of normalisation rules has a positive effect on all texts. In fact, the largest error reduction for previously unseen text, 30.7% as compared to the average 21.5%, is achieved for the church text from 1593, where the proportion of tokens with a correct modern spelling increases from 47.8% to 63.8% after normalisation. This shows that the normalisation rules are successful for texts from other centuries and genres compared to the text used for developing the rules.

This table also shows the results for part of the text *Per Larssons dombok* (1638). As noted earlier, a sample from this text was used for developing the empirically based normalisation rules. Even though a disjoint sample is used for evaluation, we see a larger error reduction for this text than for all the other texts (38.6% as compared to the average 21.5%). This indicates that the normalisation rules are somewhat biased towards the text on which the rules were based, even though the results are very promising for the other texts as well.

Development Text (Court Records)			
Text	**Orig.**	**Norm.**	**ErrRed.**
1638†	56.7%	73.4%	38.6%
Court Records			
Text	**Orig.**	**Norm.**	**ErrRed.**
1602–1605	64.6%	69.8%	14.7%
1615–1645	62.4%	71.8%	25.0%
1649–1686	64.2%	74.9%	29.9%
1649–1689	67.3%	74.7%	22.6%
1728–1741	69.2%	72.8%	11.7%
1736–1737	73.7%	78.8%	20.5%
1742–1760	67.9%	74.4%	20.2%
1761–1783	74.5%	77.1%	10.2%
1784–1795	80.8%	83.1%	19.2%
1796–1812	78.8%	81.2%	11.3%
Church Documents			
Text	**Orig.**	**Norm.**	**ErrRed.**
1527	53.5%	64.4%	23.4%
1571	51.9%	63.9%	24.9%
1593	47.8%	63.8%	30.7%
1686	64.7%	71.5%	19.3%
Average			
1527–1812	65.2%	73.0%	21.5%

Table 3: Normalisation results, given in percentage of tokens with the same spelling as in the gold standard, before and after normalisation. Orig = Proportion of words in the original text that are identical to the modern spelling. Norm = Proportion of words in the normalised text that are identical to the modern spelling. ErrRed = Error reduction. The sample taken from the same text as the (disjoint) sample used as a basis for developing the normalisation rules is marked by a † sign.

Naturally, the normalisation rules have a larger impact on the oldest texts, since these differ from the modern spelling to a larger extent, meaning that there are more words that would need normalisation. Hence, the smallest error reduction (11.3%) is observed for the youngest text (*Stora Malm*, 1796–1812), where the proportion of correctly normalised tokens increases from the original 78.8% to 81.2%. Averaging over the numbers in table 3, we see that 16th century text (1527–1593) yields an error reduction of 26.3% as compared to 22.3% for 17th century text (1602–1689)

[2] It should be noted that the part of *Per Larssons dombok* that is used for evaluation, is disjoint from the sample used for developing the normalisation rules.

and 18.6% for 18th century text (1728–1812). Still, the overall percentage of tokens in the normalised text that are identical to the gold standard spelling is generally higher for younger texts.

Interestingly, the effect of the normalisation rules seems not to be dependent only on time period and/or the proportion of original tokens that need normalisation. For example, the court records text from 1602–1605, the court records text from 1649–1686 and the church text from 1686, all have a proportion of 64-65% of words that do not need normalisation. However, for the 1649–1686 text, error reduction is twice as high (29.9%) as for the older 1602–1605 text (14.7%), and also much higher than for the church text from 1686 (19.3%). This could indicate that for texts with equal prerequisites, the normalisation rules work better for texts that are close in time and genre to the text used for developing the normalisation rules.

6.2 Tagging and Verb Extraction

The main goal of the normalisation process is to improve accuracy for language technology tools applied after normalisation, i.e. tagging and parsing. As argued in section 5, there is currently no gold standard for evaluating the automatically tagged test texts. However, all the verbs in the gold standard have been manually assigned a verb label, and we may indirectly evaluate tagging accuracy by comparing the verb tags produced by the tagger, to the words marked as verbs in the gold standard. The tagger used for this purpose is the HunPOS tagger (Halácsy et al., 2007), a free and open source reimplementation of the HMM-based TnT-tagger by Brants (2000). The tagger is used with a pre-trained language model based on the Stockholm-Umeå Corpus (SUC), a balanced, manually annotated corpus of different text types representative of the Swedish language in the 1990s, comprising approximately one million tokens (Gustafson-Capková and Hartmann, 2006).

The precision and recall measures for the verb extraction comparison are presented in table 4. The results are also compared to the results of verb identification for contemporary Swedish text from the Stockholm-Umeå Corpus. Since the tagger used in the experiments on historical texts is trained on the whole of SUC, we trained a new model for the tagger in order not to evaluate on the same data as the tagger has been trained. The tagging model used for annotating contemporary text hence includes all tokens in SUC except for the tokens reserved for evaluation.

Development Text (Court Records)				
	Orig.		**Norm.**	
Text	**Prec.**	**Rec.**	**Prec.**	**Rec.**
1638†	68.8%	51.8%	82.3%	85.5%
Court Records				
	Orig.		**Norm.**	
Text	**Prec.**	**Rec.**	**Prec.**	**Rec.**
1602–1605	72.5%	61.2%	73.1%	76.3%
1615–1645	78.8%	53.9%	78.9%	68.7%
1649–1686	74.5%	65.6%	85.2%	76.7%
1649–1689	77.2%	62.1%	83.7%	82.0%
1728–1741	84.8%	74.0%	85.1%	83.6%
1736–1737	82.1%	68.3%	85.9%	73.7%
1742–1760	85.3%	76.7%	86.0%	85.5%
1761–1783	81.7%	79.3%	85.0%	83.7%
1784–1795	90.6%	88.0%	90.4%	90.4%
1796–1812	84.4%	83.3%	85.1%	87.8%
Church Documents				
	Orig.		**Norm.**	
Text	**Prec.**	**Rec.**	**Prec.**	**Rec.**
1527	70.4%	51.7%	67.4%	70.0%
1571	72.7%	66.8%	73.2%	78.1%
1593	63.8%	45.8%	66.1%	68.4%
1686	86.7%	66.7%	84.4%	71.3%
Average				
	Orig.		**Norm.**	
1527–1812	78.3%	66.3%	80.8%	78.8%
Contemporary Text (SUC)				
Text	**Orig.**		**Norm.**	
1990s	99.1%	99.1%	–	–

Table 4: Verb extraction results after normalisation. The sample taken from the same text as the (disjoint) sample used as a basis for developing the normalisation rules is marked by a † sign.

From the verb extraction results, it is noticeable that especially recall improves to a great extent for all texts. On average, recall improves from 66.3% for raw input text to 78.8% for the normalised version of the text. This is still not very close to the 99.1% recall noted for verb identifi-

cation in the contemporary SUC sample, but high enough to be useful for our purposes. The best results are achieved for the text from 1784–1795, with a 90.4% recall.

As expected, the lowest results are generally observed for the oldest texts, even though there are exceptions. For example the church document from 1571 has a recall of 78.1%, whereas the substantially younger court document from 1736-1737 yields the lower score of 73.7%. This is remarkable also in the sense that the 1571 text has a lower recall before normalisation (66.8%) than the younger text (68.3%). One reason could be that the 1571 text is a little closer in time to the text used for generating the normalisation rules (from 1638).

In most cases, precision also improves slightly with normalisation. For the 1527 text however, precision drops from 70.4% to 67.4% when adding the normalisation rules. The leap in recall from 51.7% to 70% is however worth the slight decrease in precision. On average, precision improves from 78.3% to 80.8%.

6.3 Parsing and Complement Extraction

As explained in section 5, there is currently no gold standard for evaluating the automatically parsed test texts. However, verbal complements in the gold standard have been marked up manually. Parsing accuracy may thus be indirectly evaluated by comparing the complements assigned by the parser to the complements given in the gold standard. For this purpose, we use a string-based evaluation method first described in Pettersson et al. (2012), where all labels and brackets are removed before comparing the segments extracted from the parser to the segments extracted in the gold standard. Each extracted instance is classified as falling into one of four mutually exclusive categories:

- Fully correct complement set
- Partially correct complement set
- Incorrect complement set
- Missing complement set

A complement set is regarded as fully correct if the output string generated by the system is identical to the corresponding gold standard string. Furthermore, a complement set is regarded as partially correct if the output string generated by the system has a non-empty overlap with the corresponding gold standard string. A (non-empty) complement set is regarded as incorrect if the output string has no overlap with the gold standard string. Finally, a complement set is regarded as missing if the output string is empty but the gold standard string is not.

Complement extraction results are shown in table 5, where for each text, the results for the "raw" input text is given on top, and the results after normalisation is given at the bottom. A striking difference is that with normalisation, even though the number of fully or partially correctly extracted complements stay more or less the same, the number of incorrectly assigned complements in general drops significantly. On average, the number of incorrectly assigned complements drops from 27.4% to 22.7%. Accordingly, the proportion of verbs that are not assigned any complements at all even though the gold standard states that there should be complements, increases to a similar extent, from 23.2% o 26.2%.

Even though the proportion of correctly extracted complements do not increase much, one should bear in mind that these numbers are calculated based on the words that have been correctly identified as verbs by the tagger, meaning that the actual number of extracted complements has increased by normalisation. It is also interesting to note that for all historical texts, the proportion of fully extracted complements is higher than for the contemporary text (38.1% vs 30.3%). This may be partly explained by different annotation conventions, since the historical gold standard corpus was annotated by us, whereas the contemporary gold standard was annotated by other people. Due to this, the results may not be directly comparable, but are still an indication of parsing performance for historical versus contemporary text. It is also worth mentioning that spelling correction captures differences in surface form at a word level, improving for example tagging based on dictionaries. Grammatical differences however include changes in word order and sentence structure, i.e. differences that need to be handled by some kind of retraining of the parser.

Development Text (Court Records)				
Text	**Fully**	**Part.**	**Inc.**	**Unass.**
	32.3%	9.9%	35.4%	22.4%
1638†	34.0%	10.3%	25.3%	30.3%
Court Records				
Text	**Fully**	**Part.**	**Inc.**	**Unass.**
	41.6%	14.6%	24.8%	19.0%
1602–1605	41.8%	12.9%	21.8%	23.5%
	39.7%	12.4%	35.9%	12.0%
1615–1645	36.7%	14.2%	30.7%	18.4%
	34.3%	16.9%	27.9%	20.9%
1649–1686	33.3%	13.4%	26.4%	26.9%
	36.8%	14.2%	26.3%	22.6%
1649–1689	41.4%	11.6%	24.7%	22.3%
	34.6%	16.0%	24.7%	24.7%
1728–1741	34.4%	15.3%	23.0%	27.3%
	43.5%	16.1%	21.7%	18.7%
1736–1737	45.8%	14.6%	18.3%	21.4%
	40.7%	14.8%	24.7%	19.8%
1742–1760	42.1%	12.5%	18.5%	26.9%
	34.2%	18.0%	22.4%	25.5%
1761–1783	33.5%	19.4%	17.1%	30.0%
	39.9%	18.6%	16.4%	25.1%
1784–1795	40.4%	17.0%	16.0%	26.6%
	41.3%	17.9%	19.0%	21.7%
1796–1812	40.7%	19.6%	19.6%	20.1%
Church Documents				
Text	**Fully**	**Part.**	**Inc.**	**Unass.**
	37.3%	11.8%	33.8%	17.1%
1527	36.0%	11.3%	33.4%	19.3%
	30.1%	6.1%	34.9%	28.8%
1571	39.2%	9.0%	23.1%	28.7%
	37.0%	11.1%	25.9%	39.5%
1593	39.7%	5.0%	18.2%	37.2%
	31.8%	10.4%	27.9%	29.9%
1686	32.6%	9.8%	24.2%	33.5%
Average				
	37.0%	13.9%	27.4%	23.2%
1527-1812	38.1	13.1%	22.7%	26.2%
Contemporary Text (SUC)				
SUC	30.3%	54.2%	9.1%	6.4%

Table 5: Complement extraction results after normalisation. Results for the "raw" input text is given on top, and results after normalisation is given at the bottom. Fully = Fully identical match. Part = Partial match. Inc = Incorrect. Unass = Unassigned. The sample taken from the same text as the (disjoint) sample used as a basis for developing the normalisation rules is marked by a † sign.

7 Conclusion

In this paper, we have presented a rule-based approach to normalisation of historical text, with the aim of making NLP tools developed for modern language applicable for analysing historical text. We have shown that a relatively small set of hand-crafted normalisation rules based on one single text, had a large positive impact on the usefulness of contemporary NLP tools also for texts from other time periods and genres than the text from which the rules were developed. Our results thus show that existing NLP tools can be successfully used for analysing historical text, even for languages without access to a large amount of linguistically annotated historical data for training the tools. The fact that rules generated from one single text proved useful for other time periods and text types as well, means that we do not even need a large, balanced corpus as a basis for rule development. The choice of text used for developing the rules may however be important for success, since older texts generally contain a higher number of instances of differently spelled words. If we choose a too modern text, the rules generated may not be very useful for older texts. For the texts studied in this paper however, we saw that the same spelling variation occurs in texts from all centuries and genres (16th, 17th and 18th century court records and Church documents), only with a slightly different frequency distribution. It would be interesting to do the same kind of evaluation for other time periods and text types as well. Furthermore, we would like to evaluate the general applicability of the method by testing the same approach on texts from various time periods and in other languages. This would of course require a different set of normalisation rules, but the general method for rule development could still be the same.

Even though the presented normalisation method works well for a range of documents, it is a time-consuming and knowledge-intense manual work to formulate the normalisation rules. Future work includes more sophisticated methods for normalisation, including automatic rule generation. It would also be interesting to explore methods for improving parsing performance, where normalisation alone may not be a sufficient ap-

proach.

References

Maria Ågren, Rosemarie Fiebranz, Erik Lindberg, and Jonas Lindström. 2011. Making verbs count. The research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review*, 59(3):271–291. Forthcoming.

Gösta Bergman. 1995. *Kortfattad svensk språkhistoria*. Prisma Magnum, Stockholm, 5th edition.

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

Thorsten Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, Seattle, Washington, USA.

Nils Edling. 1937. *Uppländska domböcker*. Almqvist & Wiksells.

Sofia Gustafson-Capková and Britt Hartmann. 2006. Manual of the Stockholm Umeå Corpus version 2.0. Technical report, December.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 209–212, Prague, Czech Republic.

Csaba Oravecz, Bálint Sass, and Eszter Simon. 2010. Semi-automatic normalization of Old Hungarian codices. In *Proceedings of the ECAI Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 55–59, Faculty of Science, University of Lisbon Lisbon, Portugal, August.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2012. Parsing the Past - Identification of Verb Constructions in Historical Text. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 65–74, Avignon, France, April. Association for Computational Linguistics.

Cristina Sánchez-Marco, Gemma Boleda, and Lluís Padró. 2011. Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, OR, USA, June. Association for Computational Linguistics.