

ACTA UNIVERSITATIS UPSALIENSIS

Studia Linguistica Upsaliensia

17

Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction

Eva Pettersson



UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Ihresalen, 21-0011, Engelska Parken, Thunbergsvägen 3H, Uppsala, Saturday, 5 March 2016 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Doctor Michael Piotrowski (Leibniz Institute of European History).

Abstract

Pettersson, E. 2016. Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. *Studia Linguistica Upsaliensia* 17. 147 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9443-8.

Historical text constitutes a rich source of information for historians and other researchers in humanities. Many texts are however not available in an electronic format, and even if they are, there is a lack of NLP tools designed to handle historical text. In my thesis, I aim to provide a generic workflow for automatic linguistic analysis and information extraction from historical text, with spelling normalisation as a core component in the pipeline. In the spelling normalisation step, the historical input text is automatically normalised to a more modern spelling, enabling the use of existing taggers and parsers trained on modern language data in the succeeding linguistic analysis step. In the final information extraction step, certain linguistic structures are identified based on the annotation labels given by the NLP tools, and ranked in accordance with the specific information need expressed by the user.

An important consideration in my implementation is that the pipeline should be applicable to different languages, time periods, genres, and information needs by simply substituting the language resources used in each module. Furthermore, the reuse of existing NLP tools developed for the modern language is crucial, considering the lack of linguistically annotated historical data combined with the high variability in historical text, making it hard to train NLP tools specifically aimed at analysing historical text.

In my evaluation, I show that spelling normalisation can be a very useful technique for easy access to historical information content, even in cases where there is little (or no) annotated historical training data available. For the specific information extraction task of automatically identifying verb phrases describing work in Early Modern Swedish text, 91 out of the 100 top-ranked instances are true positives in the best setting.

Keywords: NLP for historical text, spelling normalisation, digital humanities, information extraction, character-based statistical machine translation, SMT, Levenshtein edit distance, language technology, computational linguistics

Eva Pettersson, Department of Linguistics and Philology, Box 635, Uppsala University, SE-75126 Uppsala, Sweden.

© Eva Pettersson 2016

ISSN 1652-1366

ISBN 978-91-554-9443-8

urn:nbn:se:uu:diva-269753 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-269753>)

To Christer Bergström, for being such a wonderful part of my life

Contents

1	Introduction	13
1.1	Research Questions and Contributions	15
1.2	Method Overview	16
1.3	Thesis Outline	18
1.4	Key Publications	19
2	NLP for Historical Text	21
2.1	Characteristics of Historical Text	22
2.1.1	Spelling	24
2.1.2	Vocabulary	24
2.1.3	Semantics	25
2.1.4	Morphology	26
2.1.5	Syntax	27
2.1.6	Sentence Boundaries and Sentence Length	29
2.1.7	Code-Switching	30
2.2	Adapting NLP Tools to the Historical Domain	32
2.3	Applying Modern NLP Tools to Historical Data	35
2.3.1	Applying Tools Off-the-Shelf	36
2.3.2	Dictionary-based Adaptation	37
2.3.3	Levenshtein-based Spelling Normalisation	39
2.3.4	Spelling Normalisation based on Phonetic Similarity ..	41
2.3.5	SMT-based Spelling Normalisation	42
2.4	Relation to Spelling Variation in Modern Data	43
	Part I: Spelling Normalisation of Historical Text	45
3	Approaches to Spelling Normalisation	47
3.1	The Gender and Work Corpus	49
3.2	Rule-based Normalisation	51
3.2.1	Spelling in Different Time Periods and Genres	52
3.2.2	Evaluation	54
3.3	Levenshtein-based Normalisation	56
3.3.1	String Length Restrictions	58
3.3.2	Edit Distance Restrictions	59
3.3.3	Weighted Edit Distance	61
3.3.4	Compound Splitting	63
3.4	Memory-based Normalisation	64

3.5	SMT-based Normalisation	65
3.5.1	Different Techniques for Word Alignment	71
3.5.2	Different Sizes of the Parallel Corpus	73
3.5.3	Different Sizes and Genres of the Target Language Corpus	74
3.5.4	Different Tools and Settings for Character Alignment ..	75
4	Multilingual Evaluation	77
4.1	Experimental Setup	77
4.1.1	English	79
4.1.2	German	79
4.1.3	Hungarian	80
4.1.4	Icelandic	80
4.1.5	Swedish	81
4.2	Results	81
5	Part I: Summary and Conclusion	87
	Part II: Linguistic Analysis of Historical Text	89
6	Verb Phrase Extraction	91
6.1	Verb Identification	93
6.2	Complement Extraction	95
7	Valencies for Improved Verb Phrase Extraction	98
7.1	Data	99
7.2	Deletion of Improbable Complements	102
7.3	Insertion of Probable Complements	102
7.4	Model Selection for Deletion	104
7.5	Model Selection for Insertion	107
7.6	Evaluation	109
8	Part II: Summary and Conclusion	111
	Part III: Information Extraction from Historical Text	113
9	Verb Phrase Ranking	115
9.1	Data	116
9.2	Conditional Probability	118
9.3	Log Likelihood Ratio	118
9.4	Bag-of-Words Classification	120
10	Evaluation	122
10.1	Evaluation Metrics	123
10.2	Conditional Probability Results	123
10.3	Log Likelihood Ratio Results	125
10.4	Bag-of-Words Classification Results	126

10.5	Summary of the Results	127
11	Part III: Summary and Conclusion	128
12	Conclusion	129
12.1	Spelling Normalisation	130
12.2	Linguistic Analysis	132
12.3	Information Extraction	133
12.4	Summary of the Results	134
12.5	Future Work	135
	References	139
	Appendix A: Early Modern Swedish Normalisation Rules	145

Acknowledgements

First of all, I am very grateful to my supervisors, Joakim Nivre and Beáta Megyesi, for their never-ending encouragement, guidance and inspiration. I would also like to thank Anna Ságvall Hein for introducing me to the area of language technology back in the 90s, and for eventually introducing me to the language technology aspects of the Gender and Work project, which became the starting point for this thesis. Special thanks is due to Jonas Lindström at the Department of History, for providing me with historical material and for interesting discussions and suggestions for research questions to solve. Credit is also due to Lasse Mårtensson, Eszter Simon, Eiríkur Rögnvaldsson, Sigrún Helgadóttir and Merja Kytö for generously sharing data to my experiments. I am very grateful to Jörg Tiedemann and Christian Hardmeier for guiding me through the SMT jungle, and to Sara Stymne for sharing with me her knowledge and technology for compound splitting. I would also like to thank Per Starbäck for technical support. Special thanks also to Pelle Weijnitz for supporting me throughout the years, both technology-wise and friendship-wise. I would also like to express my gratitude to my opponent at the mock defense, Gerlof Bouma, for valuable comments and suggestions for improvement. Furthermore, I am very grateful to all the people at the Department of Linguistics and Philology at Uppsala University in general, and to the computational linguistics group in particular, for contributing to making work such a nice place to go to. Special thanks is also due to the knitting group at the Department; Helena Löthman, Jenny Rahbek, Ina Sörlid and Therese Tiedemann. When I needed to relax from thesis work, you inspired me to disappear into the wonderful world of knitting. Last, but certainly not least, my deepest thanks to Christer Bergström for being such a wonderful part of my life. I love you to the moon and back!

1. Introduction

Natural Language Processing (NLP) of historical text is of great interest not only to the language technology community, but also for the preservation of our cultural heritage and as an aid for researchers in humanities and social sciences working with historical material. Language technology could contribute to both these areas in the form of OCR techniques for digitisation of old texts, information extraction techniques for making the information concealed in the texts searchable and easily accessible, and linguistic analysis tools to make more sophisticated analyses of the texts possible.

Today, a large proportion of the historical material is still handwritten, or possibly printed, and not electronically available in a searchable format. This often means that researchers working with old texts are left with the only option of manually going through large volumes of text in search for relevant information; a time-consuming work. Even in cases where the text has been digitised, contemporary tools for linguistic analysis and information extraction are generally not suitable for processing the text, since historical text differs in many aspects from modern text.

One of the most striking differences concerns spelling. Not only are words in historical documents spelled differently than today. The lack of spelling conventions in historical times also leads to a substantial variation in spelling between documents written within different genres or by different authors, and even within the same text written by the same author. This is problematic, since language technology tools often to some degree rely on frequencies of words and word sequences. Due to the absence of writing conventions in general, historical text also often exhibits varying word order and inconsistent use of punctuation. Other common obstructive features are longer sentences, a somewhat different vocabulary, and a more complex morphology.

A vast majority of the available NLP tools, such as taggers for morphological analysis and parsers for syntactic analysis, are trained on contemporary language. As argued, the characteristics of historical text however pose problems when applying modern NLP tools to historical data. One possible solution to this would be domain adaptation by developing new language technology tools specifically aimed at handling historical text. Still, the inconsistencies in spelling and syntax would be problematic for this solution as well, due to data sparseness issues during training, leading to a high degree of unknown word forms and sentence structures when applying the models to previously unseen text. In addition, historical language is often under-resourced with regard to annotated data needed for training NLP tools. This problem is further

aggravated by the fact that the concept of historical text may refer to texts from a long period of time, during which language has changed. NLP tools trained on 13th century texts may thus not perform well on texts from the 18th century.

Even though there are differences between old and modern text, there are also similarities, and a native speaker of the language is often able to understand all or part of an old text in spite of the odd spelling and structure (with varying success due to aspects such as the age and genre of the text). Bearing this in mind, one way to get around the problem of adapting NLP tools to historical text is to instead adapt the historical text to the contemporary NLP tools by modernising the text. This could be done in several ways, depending on what features should be handled (e.g. orthography, vocabulary, morphology, syntax, and/or punctuation). The most prominent difference, with a high impact on NLP performance, is however spelling. In this thesis, I therefore add spelling normalisation as a core component for linguistic analysis and information extraction from historical text. The hypothesis is that after spelling normalisation, state-of-the-art taggers and parsers developed for the modern language can be successfully applied to the old text in its modernised spelling, with the aim of performing the linguistic analysis needed to extract the desired information from the text. In this context, information extraction is defined as presenting to the user a ranked list of candidate phrases that have a high probability of containing the requested information. The whole process is illustrated in Figure 1.1.

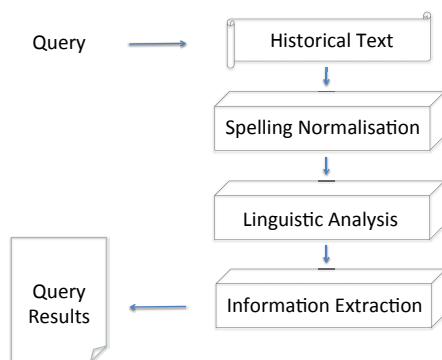


Figure 1.1. Overview of my approach to information extraction from historical text.

As regards the information extraction step, a specific information need was expressed by the researchers at the Department of History at Uppsala University within the framework of the *Gender and Work* project. In this project, historians are storing information in a database on what men and women did for a living in the Early Modern Swedish society (i.e. approximately 1550–1800). During this work they have found that working activities in their source ma-

terial are most often expressed in the form of verb phrases, such as *to fish herring* or *to sell clothes* [Ågren et al., 2011]. The information extraction step in this thesis is thus focused on the extraction of verb phrases describing work from Early Modern Swedish text. The central spelling normalisation step, on the other hand, is evaluated for several languages, time periods and genres.

1.1 Research Questions and Contributions

In this thesis, I focus on the use of contemporary NLP tools for linguistic analysis and information extraction from historical text. As discussed above, NLP tools available for the modern language generally do not perform well when applied to historical text. However, the characteristics of historical text, combined with the lack of linguistically annotated historical data, also make it difficult to adapt modern NLP tools to the domain of historical text. Nevertheless, automatic linguistic analysis of old texts would enable more sophisticated information extraction techniques than key word search, and thus contribute to making old material more easily accessible and searchable. My hypothesis is that automatic spelling normalisation can enable contemporary NLP tools to be successfully applied to historical text for the purpose of information extraction. The following research questions are in focus:

1. How successful are different approaches to spelling normalisation of historical text?
2. To what extent can spelling normalisation enable the use of modern NLP tools for linguistic analysis of historical text?
3. To what extent can spelling normalisation and subsequent linguistic analysis enable extraction of relevant information from historical text?
4. How can information extracted from historical text be ranked, so that the most relevant instances are presented at the top of the results list?

An important consideration in the implementation of my information extraction pipeline is that all parts of the chain should be generic, and not tailored towards a single source language, time period, document type, or information need. Due to the general lack of annotated historical data, it is also desirable to explore methods that do not rely on large amounts of annotated training data. A further aim is that linguistic analysis of historical text should be possible for any language where there is a basic set of language resources and tools available for the modern version of the language. Hence, the following constraints apply:

1. The methods used for spelling normalisation should be generally applicable to different source languages, time periods, and genres.

2. The information extraction step should not be dependent on a full-fledged information extraction system tailored towards a specific language or information need.
3. The NLP pipeline should be applicable also in cases where only small amounts of annotated historical data are available.
4. NLP tools available for the modern language are to be reused in the linguistic analysis phase.

The main contributions of the thesis are:

1. The development and evaluation of a generic pipeline for information extraction from historical text, based on spelling normalisation combined with state-of-the-art NLP tools available for the modern language.
2. The development and evaluation of different approaches to spelling normalisation of historical text.
3. The development and evaluation of techniques for applying contemporary linguistic analysis tools to historical text.
4. The development and evaluation of different approaches to ranked information retrieval from historical text.

Each step is evaluated both regarding performance (e.g. in terms of precision and recall measures) and by how well the chosen methods comply with the criteria of being generic and not requiring large amounts of annotated historical training data.

1.2 Method Overview

My approach to information extraction from historical text includes several alternative methods for spelling normalisation, succeeded by linguistic analysis using modern state-of-the-art tools, information extraction based on the linguistically annotated data, and ranking of relevant phrases in the extracted data. An overview is given in Figure 1.2.

The first step is tokenisation of the historical source text. Since I aim for a pipeline where contemporary NLP tools are to be used without domain adaptation, standard tokenisation methods for the modern language are applied to the source text, disregarding the fact that historical text may be structurally different from modern text, and may also contain abbreviations that are not part of the modern written language.

The next step is normalisation, where the tokenised text is transformed to a more modern spelling, using one of four methods: 1) rule-based normalisation, 2) memory-based normalisation, 3) Levenshtein-based normalisation,

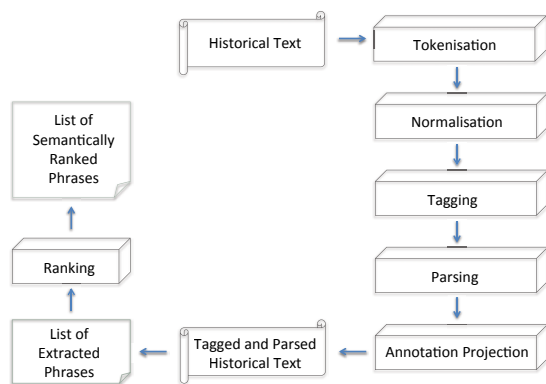


Figure 1.2. Method overview.

or 4) SMT-based normalisation. All four methods are mainly based on empirical findings on how spelling in old texts corresponds to spelling in modern texts (although known spelling changes are also taken into account in the rule-based method). It could be argued that part of the historical linguistics research has focused on diachronic changes in spelling for certain languages, and that the outcome of this research should be incorporated in the normalisation process. Since I aim at developing generic, language-independent methods for spelling normalisation, empirically based normalisation approaches are however favoured, even though the performance of these methods could probably be improved by also adding language-specific information based on historical linguistics research.

After normalisation, the text is used as input for tagging and parsing, performed by available tools trained for the modern language. The annotations given by the tagger and the parser are then projected back to the text in its original spelling, resulting in a tagged and parsed version of the historical text, from which phrases of interest are extracted based on the annotation labels. In the ranking phase, the extracted phrases are ordered so that relevant phrases are presented at the top of the results list. For information extraction and ranking, the Gender and Work project (presented above) provides both an interesting information need (verb phrases describing work) and a corpus suitable for training and evaluation (the Gender and Work database). The information extraction step is therefore geared towards the extraction of verb phrases from Early Modern Swedish text, with verb phrases that are likely to describe work presented at the top of the results list. The implemented extraction and ranking methods are however of a general nature, with the aim of making them easily applicable to other information needs as well. No full-

fledged information extraction system is required in the pipeline, and shifting to other information needs is done by simply altering the training data.

1.3 Thesis Outline

The outline of the thesis is as follows. In Chapter 2, I start by giving an overview of some characteristic features of historical language that are of importance in the context of NLP for historical text. Thereafter, I present some of the research that has been conducted within the area of NLP for historical text, describing different approaches to handling language variation and other characteristics of historical text, some of which have inspired me in the development of my own methods for spelling normalisation. The rest of the thesis is divided into three main parts:

The first part, *Spelling Normalisation*, begins with Chapter 3, where my four approaches to spelling modernisation of historical text are described in detail and evaluated one by one. The evaluation of each method is mainly focused on exploring optimal settings and/or how well the method complies with the criteria of being applicable to different time periods and genres, and in cases where no (or only small amounts of) training data are available. In Chapter 4, the different normalisation techniques are evaluated in comparison to each other when applied to different source languages. Conclusions are drawn in Chapter 5.

The second part, *Linguistic Analysis of Historical Text*, focuses on the task of applying modern NLP tools to automatically normalised historical text. In Chapter 6, I describe my experiments on applying a modern Swedish tagger and parser to Early Modern Swedish text, for the task of automatic verb phrase extraction. The verb phrase extraction task is further divided into *verb identification*, based on the words analysed as verbs by the tagger, and *complement extraction*, in which certain phrases are extracted as complements to the verbs, based on the annotation labels given by the parser. In Chapter 7, I describe a method for improving the complement extraction part by adding verb valency information in a post-processing step. In this step, complements suggested by the annotation labels are deleted if not compatible with the information listed for the head verb in the valency resources. Likewise, complements suggested by the valency frame, but not found by the parser, are inserted on the basis of phrases occurring in the near context of the head verb. Part II is summarised, and conclusions are drawn, in Chapter 8.

In the third part, *Information Extraction from Historical Text*, the actual information extraction task is in focus. In Chapter 9 different methods for ranking relevant verb phrases are presented, with verb phrases describing work as a case study. The results are given in Chapter 10, with comparisons between the different methods. Conclusions are drawn in Chapter 11.

Finally, the thesis is summarised in Chapter 12, with conclusions, suggestions for future research and some final remarks.

1.4 Key Publications

Parts of the results presented in this thesis have been published in earlier work by the author, as listed below:

- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2015)
Ranking Relevant Verb Phrases Extracted from Historical Text. In: Proceedings of the 9th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 39–47.
- Eva Pettersson and Joakim Nivre (2015)
Improving Verb Phrase Extraction from Historical Text by use of Verb Valency Frames. In: Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA), pages 153–162.
- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2014)
Verb Phrase Extraction in a Historical Context. The First Swedish National SWE-CLARIN Workshop at the Swedish Language Technology Conference (SLTC). Uppsala, Sweden.
- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2014)
A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text. In: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 32–41.
- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2013)
Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA), pages 163–179.
- Eva Pettersson, Beáta Megyesi and Jörg Tiedemann (2013)
An SMT Approach to Automatic Annotation of Historical Text. In: Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA, pages 54–69.
- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2012)
Rule-Based Normalisation of Historical Text - A Diachronic Study. In: Proceedings of the First International Workshop on Language Technology for Historical Text(s), pages 333–341.
- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2012)
Parsing the Past - Identification of Verb Constructions in Historical Text.

In: Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 65–74.

- Eva Pettersson and Joakim Nivre (2011)
Automatic Verb Extraction from Historical Swedish Texts. In: Proceedings of the 5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 87–95.

2. NLP for Historical Text

In recent years, there has been a growing interest in the area of digital humanities and NLP for historical text. One indication of this is the emergence of workshops specifically focusing on this field, such as the workshop series on *Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH, from 2007 and onwards), the workshop on *Computational Historical Linguistics* (in conjunction with the NODALIDA conference 2013), and the workshop on *Language Resources and Technologies for Processing and Linking Historical Documents and Archives* (LRT4HDA, in conjunction with the LREC conference 2014). Furthermore, a number of digital humanities centres have been established around the world, bringing together researchers across disciplines to collaborate on the issue of making humanities research material digitally accessible and explorable. Some examples are the *Digital Humanities Centre* at the University of Nottingham (UK), the *Göttingen Centre for Digital Humanities* (GCDH, Germany), the *Center for Digital Humanities* at Princeton University (CDH, USA), the *Ghent Center for Digital Humanities* (GhentCDH, Belgium), and the *Centre for Digital Humanities* at the University of Gothenburg (Sweden).

For the specific subfield of NLP for historical text, the concept of ‘historical text’ needs to be clarified. Historical linguists categorise languages into different time periods, based on evolutionary stages of the language. For example, the English language is divided into Old English (~450–1150), Middle English (~1150–1500), Early Modern English (~1500–1700), and Modern English (~1700–) [van Gelderen, 2006]. As Piotrowski [2012] points out, these categorisations may however not be appropriate to use in the context of NLP for historical text. That is because texts produced within the time period classified as ‘modern’ by linguists (e.g. 18th century English) may still exhibit differences in for example spelling and vocabulary as compared to present-day texts, making it hard to achieve state-of-the-art results when applying contemporary NLP tools to the text. In the context of NLP for historical text, I therefore adopt the definition suggested by Piotrowski [2012], where historical language is more vaguely defined as older versions of a language where the differences as compared to modern language pose problems when applying contemporary NLP tools to the text.

The field of NLP for historical text is still to a large extent unexplored. Due to the recent digitisation of large volumes of historical material, there is however an increasing need for automatic analysis of historical text, and there have been several attempts to provide solutions for this. Depending on the

resources available and the primary goal of the analysis, two main approaches can be distinguished:

1. Linguistic analysis of historical text using NLP tools specifically adapted to the task of analysing historical text.
2. Linguistic analysis of historical text using NLP tools developed for the modern language, with or without adaptation of the input data.

In the following, I start by describing some characteristics of historical text, with the aim of providing a better understanding of the issues that need to be handled in the context of NLP for historical text. Thereafter, I present an overview of previous approaches to automatic linguistic analysis of historical text, with a few representative examples of research conducted for each approach presented.

2.1 Characteristics of Historical Text

There are a number of characteristics of historical written language to be addressed in the context of NLP for historical text. Considering that the concept of historical text is wide, covering a range of different time periods, genres, and languages, the characteristics vary between different texts. Typically, the older the text is, the more different it is in comparison to contemporary text. There are also differences between languages. For example, the Icelandic language has not changed much during the past thousand years, neither in morphology nor in syntax. Two reasonable explanations are the strong literary tradition on Iceland, and the isolated location of the country, resulting in limited contact with other languages that could influence the Icelandic language. Rögnvaldsson et al. [2012] state that:

“...the inflectional system and the morphology has in all relevant respects remained unchanged from the earliest texts up to the present, although a number of nouns have shifted inflectional class, a few strong verbs have become weak, one inflectional class of nouns has been lost, and the dual in personal and possessive pronoun has disappeared. The syntax is also basically the same, although a number of changes have occurred. The changes mainly involve word order, especially within the verb phrase, and the development of new modal constructions.” [Rögnvaldsson et al., 2012]

For other languages, the changes are more profound. The English language, for instance, has long since had intense contact with other languages, leading to major changes in the language throughout history. Due to the Viking invasions in the 8th to 10th centuries, the Scandinavian languages have had a large impact on English grammar and vocabulary. It has also been estimated

that approximately half of the English vocabulary originates from French and Latin [van Gelderen, 2006].

As exemplified above, language change is not a uniform process where the same changes occur in all languages. Nevertheless, historical text generally differ in orthography, vocabulary, semantics, morphology, and syntax, as compared to present-day standard language. Due to the lack of spelling conventions in the past, historical text typically shows a high degree of spelling variance, both between different texts and within the same document (see further Section 2.1.1). Concerning vocabulary, contact with speakers of other languages often leads to borrowing of words and concepts from the other language(s). Furthermore, as a result of technical and societal development, some terms become rare over time and eventually disappear from the language. Thus, historical text may contain archaic word forms that are unfamiliar to the present-day reader (see further Section 2.1.2). Similarly, semantic change is a frequently occurring phenomenon in language history. Common types of semantic change are widening and narrowing of word senses, as well as metaphorical use of certain word forms to describe other concepts than the originally intended one. As a result of these shifts in meaning, the present-day reader may find it hard to interpret the meaning of a historical sentence, even in cases where the actual words used in the sentence are familiar to the reader (see further Section 2.1.3).

Morphological change is another source of differences between historical and modern text. Due to the process of morphological levelling, the endings in historical language are generally more irregular and complex (see further Section 2.1.4). Regarding syntactic change, a common difference in historical text as compared to modern text is word order. Differences in word order are often related to morphological change. For languages such as English, that have developed from a highly inflective language into a more analytic language, word order has become more important for the interpretation of a sentence (see further Section 2.1.5). For such languages, the word order is therefore typically more strict in present-day language than in historical language [Campbell, 2013, Lehmann, 1992, Ringe and Eska, 2013].

At the sentence level, it is also worth noting that sentences in historical text may differ in length as compared to sentences in contemporary standard text (see further Section 2.1.6). It is also often hard to distinguish sentence boundaries, due to an inconsistent use of punctuation [Bouma and Adesam, 2013]. Finally, code-switching is a much more common phenomenon in historical text than in contemporary text. Thus, different languages are often mixed within the same historical document, or even within the same sentence (see further Section 2.1.7).

In the following, I describe the various levels of differences between historical and modern text in more detail, with a few illustrative examples mainly from English, German, and Swedish.

2.1.1 Spelling

One of the most striking differences between historical and modern text concerns spelling. Piotrowski [2012] makes a distinction between *diachronic* and *synchronic* spelling variance in historical text. Diachronic spelling differences are due to the fact that language changes over time, resulting in differences in spelling between historical text and modern text. Examples of diachronic spelling differences in Early Modern Swedish (~1526–1732) described in Bergman [1995] are:

- The duplication of long vowels, as in:

Early Modern Swedish	Contemporary Swedish	
<i>saak</i>	<i>sak</i>	‘thing’
<i>stoor</i>	<i>stor</i>	‘big/large’

- The use of *fv* instead of the present-day *v*, as in:

Early Modern Swedish	Contemporary Swedish	
<i>öfver</i>	<i>över</i>	‘over’

- The use of *gh* and *dh* instead of the present-day *g* and *d*, as in:

Early Modern Swedish	Contemporary Swedish	
<i>någhon</i>	<i>någon</i>	‘somebody’
<i>fadhren</i>	<i>fadern</i>	‘the father’

The concept of synchronic spelling variance on the other hand, refers to differences in spelling in texts produced within the same time period. These inconsistent spellings are due to the lack of spelling conventions in historical time, causing writers to spell the words the way they sound, which obviously is a subjective assessment. The lack of spelling conventions typically leads to spelling variance between different texts, but there are also numerous examples of varying spelling within the same document. One example given by Bergman [1995] is the Swedish book of prayers *Svenska tideboken* printed in 1525, where the pronoun *mig* (‘me’ or ‘myself’) is spelled in six different ways throughout the book: *mig*, *migh*, *mik*, *mic*, *mich*, and *mech*.

2.1.2 Vocabulary

Concerning vocabulary, new words are constantly entering languages, and old words are becoming less frequent or eventually non-existing. In some cases, the need for a new word is invoked by technological development. In the absence of appropriate terms to use for the new concept, a word is then often borrowed from another language. Thus, many languages have similar terms for describing technological innovations. For example, the word for *car* is *avtomobil* in Russian, *auto* in Finnish, and *bil* in Swedish.

Loan words may also enter a language for prestigious reasons. The English language has borrowed a considerable amount of food terms from French, such as *pork* (from French *porc*) and *beef* (from French *bœuf*), mainly due to the reason that French had a higher social status than English during the Norman French dominance of the country (1066–1300) [Campbell, 2013].

The examples given so far are all loan words that are still in use in the modern language. Sometimes however, new words are only in use for a limited period of time, and do not survive to the modern language. van Gelderen [2006] lists a set of words that were introduced to the English language during the Early Modern period (1500–1700), that are not in use anymore, such as:

Early Modern English	Modern English Equivalent
<i>adminiculation</i>	<i>aid</i>
<i>anacephalize</i>	<i>to summarise</i>
<i>eximious</i>	<i>excellent</i>

Likewise for German, Salmons [2012] lists Early New High German (1350–1650) words that are no longer in use, such as:

Early New High German	Modern German	
<i>liberei/librari</i>	<i>Bibliothek</i>	‘library’
<i>triangel</i>	<i>Dreieck</i>	‘triangle’
<i>akkord</i>	<i>Vertrag</i>	‘treaty’

Similar to the fact that new words enter the language due to technological development or changes in societal structure, old word forms may vanish for the same reason. One example is the word form *brofogde*, which is a Swedish occupational title for someone who was employed for surveying bridges. This word is not part of the contemporary Swedish vocabulary, since this kind of occupation does not exist anymore. Other word forms simply get out of fashion and become less frequent or extinct. Examples of archaic word forms that are rare or non-existing in contemporary Swedish text are *absentera* (old word for ‘be absent’), *umgälla* (old word for ‘suffer for’), and *ärna* (old word for ‘intend to’).

2.1.3 Semantics

Other common word-level differences in historical text are due to semantic change. The fact that the meaning of words tend to change over time can lead to misinterpretations for present-day humans (or machines) when analysing older texts. Below are a few examples of a shift in meaning from Early Modern Swedish to contemporary Swedish, as given by Bergman [1995]:

Word Form	Early Modern Meaning	Contemporary Meaning
<i>artig</i>	‘beautiful’	‘polite’
<i>rolig</i>	‘calm’	‘funny’
<i>snäll</i>	‘fast’	‘kind’

There are different ways to classify the types of semantic change that may take place in a language throughout history. Three of the classes commonly referred to in the field of historical linguistics are (as defined by [Campbell, 2013]):

- **Widening** (generalisation/extension/broadening)

The term widening refers to the process in which the meaning of a word is generalised to cover more instances than in its original use. A Spanish example is the word *pájaro*, which originally meant ‘sparrow’, but has shifted in meaning to the more general ‘bird’. An English example is the word *dog*, which originally referred to a specific breed of dog only.

- **Narrowing** (specialisation/restriction)

Narrowing is the opposite process of widening. In narrowing, the meaning of a word is restricted to cover only a subset of the instances included in the original definition of the word. Two English examples are *meat* (originally meaning ‘food’ in general), and *deer* (originally meaning ‘animal’ in general).

- **Metaphorical Extension**

In the process of metaphorical extension, the meaning of a word is extended to cover instances that are similar in some way to the original definition of the word. van Gelderen [2006] exemplifies this phenomenon by the English word form *crane* which originally is the name of a bird, but is also used for a mechanical device that resembles the shape of a bird.

2.1.4 Morphology

Apart from differences in spelling, vocabulary, and semantics, morphological changes are also common in language history. One frequently occurring phenomenon is referred to in historical linguistics literature as *analogical levelling* [Campbell, 2013]. In the process of levelling, the number of morphological forms are reduced, resulting in more uniform inflectional paradigms. In the typical case, uncommon or irregular inflectional patterns are generalised into a more common pattern. Recent examples in the English language are the shift in inflection from a strong verb paradigm to a weak verb paradigm for the verbs *cleave* and *strive*:

Historical English Inflection	Modern English Inflection
<i>cleave - clove - cloven</i>	<i>cleave - cleaved - cleaved</i>
<i>strive - strove - striven</i>	<i>strive - strived - strived</i>

Similar levelling has occurred also for adjectival comparative forms, as in the following example:

Historical English Inflection	Modern English Inflection
<i>old - elder - eldest</i>	<i>old - older - oldest</i>

For German, many verbs with vowel alternation in singular and plural form have changed into having the same vowel in both forms. Campbell [2013] exemplifies this by comparing the writing of Martin Luther (1483–1546) to Modern German inflection:

Martin Luther	Modern German	
<i>er bleyb/sie blieben</i>	<i>er blieb/sie blieben</i>	‘he stayed/they stayed’
<i>er fand/sie funden</i>	<i>er fand/sie fanden</i>	‘he found/they found’

For the Swedish language, verbs used to be inflected for number, whereas from the first part of the 20th century and onwards, the same verb form is used regardless of number. Hence, older Swedish texts contain plural verb forms that are no longer in use in present-day Swedish, such as *gingo* (plural ‘went’), *sågho* (plural ‘saw’), and *äro* (plural ‘were’). Likewise, older versions of Swedish make a distinction between nouns in the nominative form and nouns in the accusative form. For example, the nominative noun forms *qwinna* (‘woman’) and *bonde* (‘farmer’) were contrasted with the accusative noun forms *qwinno* and *bonda* respectively. In contemporary Swedish, no distinction is made between nominative and accusative noun forms, and the nominative form is used in both contexts [Bergman, 1995].

2.1.5 Syntax

Regarding syntax, a frequently observed difference between old and modern versions of a language concerns word order. Old English, for example, had a much freer word order than Modern English. This is due to a major change in the history of the English language, during which English went from being a *synthetic language* to being a (mostly) *analytic language*. Synthetic languages are highly inflected languages, where word endings are used to indicate grammatical functions, making word order less important for the interpretation of a sentence. Thus, the word order is often less strict in synthetic languages. Analytic languages, on the other hand, have fewer word endings, making word order an important clue for interpreting the grammatical functions of the words in a sentence [van Gelderen, 2006]. Apart from word order, *grammaticalisation* is another important feature in the process of becoming an analytic language. Grammaticalisation is a term used for describing the process in which

previously independent word forms weaken in meaning and develop into an auxiliary or a grammatical marker [Campbell, 2013]. This typically leads to the introduction of auxiliary verbs and prepositions. For example, the English word form *to* was used in Old English as a content word with a specific locational meaning. In Modern English, this word has however turned into an object marker in the form of a preposition. The English shift from synthetic to analytic language is described by van Gelderen [2006] as follows:

“Comparing the many endings and few words of Old and Modern English, we see that the main change between the two stages is that of a language with free word order and many endings but no ‘small’ words such as **the** or **to** becoming a language with strict word order, few endings and many ‘small’ words.” [van Gelderen, 2006]

The Swedish language also had a freer word order in the past. Old Swedish allowed for inverted word order in main clauses, that is main clauses with an initial finite verb, as in the following segment from *Upplandslagen* written in 1296: *Liggær lik a wighwalli* (‘Is corpse on manslaughter site’). This would translate into the contemporary Swedish verb-second word order *Det ligger ett lik på dråpsplatsen* (‘There is a corpse on the manslaughter site’).

The freer word order in older versions of Swedish also allowed for finite verbs to be placed at the end of subordinate clauses, which is not grammatical in contemporary Swedish. Thus an Early Modern Swedish phrase like *om man i hächtelse sätter* (‘if one in custody **is**’) would correspond to the contemporary Swedish word order *om man sätter i häkte* (‘if one **is** in custody’).

Another syntactic change in the Swedish language concerns verb-particle constructions. Many verb-particle constructions where the verb and its particle were previously written as separate words, have merged into a single word in contemporary Swedish. One example is the verb *angripa* (‘to attack’), which was previously realised as two separate word forms; the particle *an/ann* and the verb *gripa* (‘to fetch’), as illustrated in the following example from Bergman [1995]:

Early Modern Swedish	Contemporary Swedish	
<i>at gripa fienden ann</i>	<i>att angripa fienden</i>	‘to attack the enemy’

The opposite also occurs, that is, that a verb-particle construction where the verb and its particle were previously merged are written as separate words in contemporary Swedish. One example from *Hammerdals tingslag* (court records text written in the period 1649-1686) is the verb *tillhålla* and its modern language equivalent *hålla till* (‘stay/reside’), as illustrated below:

Early Modern Swedish	Contemporary Swedish	
<i>uthi gården tillhållit</i>	<i>uti gården hållit till</i>	‘at the farm resided ’

2.1.6 Sentence Boundaries and Sentence Length

One key factor for successful linguistic analysis of a text is a correct sentence segmentation, since most NLP tools such as taggers and parsers are applied one sentence at a time. For standard modern text, this is a more or less solved problem, at least for languages where sentence boundaries are marked by punctuation followed by an uppercase letter. In historical text however, sentence boundaries are often marked in a more inconsistent way. Bouma and Adesam [2013] found that in Old Swedish texts (13th to 16th century) it is often hard to determine where one sentence ends and another sentence begins. In some contexts, the sentence boundary is marked by a full stop followed by an uppercase letter, as in contemporary Swedish text. It is however also common to use characters such as slashes, commas, or semi-colons as delimiters, succeeded either by an uppercase letter or a lowercase letter. Sometimes an uppercase letter in itself signals a sentence boundary, without any preceding punctuation. There are also cases where boundaries are totally unmarked. In the following example from Bouma and Adesam [2013], the boundary marking strategy varies between slash and no marking at all (where || have been inserted to illustrate sentence boundaries not marked in the original version of the text):

sua ma han þem æpte sinum vilia bøggha / at þe mago kallas oc vara / þe hælgo kirkiu dorakroka / || mz þe hælgho kirkio gulfe / menas biskopa / oc værulsleke klærka / || þera giri ær sua diup / þæt þær ma ingte i grynna / oc þera høgghærfe oc skøro lifærne gar af fragh [...] || þætta ma pafin an han vil mykyt at bættra || rapæ allum biskopum þy følghia i gozs oc allum andrum þingum sum þu hørþe at honum var rapæt siælfum at gørra

‘Thus he may bend them after his will, so that they may [truly] be called the holy church’s door hinges. With the holy church’s floor, the bishops are meant and secular clergy. Their greed is so deep that nothing can reach its bottom, and their vanity and sinfulness is infamous [...]. This the pope can make a lot better, if he wants. Advise all bishops to do in questions of property and all other things like you heard he himself was advised to do.’
(*Brigitta-autograferna*)

Longer sentences are typically harder to analyse automatically. This is problematic in the context of NLP for historical text, since the inconsistent marking of sentence boundaries in these texts may cause sentence segmentation problems, leading to longer chunks for the NLP tools to handle. There is also some research indicating that, apart from sentence segmentation issues, the writing style in older times also favoured longer sentences than in modern text. Due to the inconsistent marking of sentence boundaries, it is difficult to perform large-scale studies on sentence length for different time periods in the past. There are however studies on sentence length for younger time periods,

during which sentence boundaries are most commonly marked by a full stop, enabling automatic calculations of sentence length. In Bergs [2012], a comparison of English newspaper text from different time periods shows that in newspapers from 1691, the average sentence length is 41.77 words, whereas a hundred years later, the average sentence length has decreased to 28.61 words per sentence. In general, the sentence length in their newspaper text collection has decreased from 35 words per sentence in 1700 to less than 20 words per sentence in 2000.

The difference in sentence length is also noticeable in the Gender and Work corpus of court records and church documents from the Early Modern Swedish period. The average sentence length in the randomly sampled part of this corpus that I use in my experiments throughout the thesis is 51 tokens, as compared to 16 tokens in the Stockholm-Umeå corpus of contemporary standard Swedish text [Ejerhed and Källgren, 1997]. This could however possibly be a genre-specific property rather than an age-specific property.

2.1.7 Code-Switching

A linguistic phenomenon arising from language contact is *code-switching*. In code-switching, a speaker (or a writer) mixes two or more languages in the same utterance (or text). There are many factors that may invoke code-switching. For example, code-switching may be used to signal social status and identity. Furthermore, for people living in bilingual communities, it is sometimes easier to find the correct term for certain concepts in another language. Gardner-Chloros [2009] provides an example from Strasbourg, the capital of Alsace in the easternmost region of France. This region has alternated between German and French dominance, resulting in a particular Alemannic dialect called Alsatian. Since 1945, however, the language of state, education and media is French, and it is common for Alsatian speakers to use French terms for concepts adhering to these areas, as illustrated in the following example, where code-switching to French is marked in **bold**:

*E **promesse de vente** isch unterschriwwe, nitt, awwer de **contrat de vente** nitt, denn ... der het's kauft vor zehn Johr, am achzehnte **janvier**; s'il le **revend avant**, hett'r e **plus-value**.*

‘The **exchange of contracts** has been signed, you see, but the **sales contract** hasn't, because ... he bought it ten years ago, on the eighteenth of **January**; if he sells before (**ten years are up**), he'll have to pay **gains tax**.’

In historical text, code-switching is a much more frequent and accepted phenomenon also in written text, both in official and non-official text. Schendl and Wright [2011] concludes that “*During the medieval period, written code-*

switching ... was both a normal phenomenon within the British Isles, and Europe-wide". Medieval code-switching occurs in a variety of genres, such as sermons, legal documents, and medical texts. The following example, given by Schendl [2002], shows code-switching between English and French in a letter from the Dean of Windsor, Richard Kingston, to King Henry IV, written in 1403 (English in **bold**):

*Please a vostre tresgraciouse Seignourie entendre que a-jourduy apres noone... qu'ils furent venuz deinz nostre **countie** pluis de .cccc. des les rebelz de Owyne, Glyn, Talgard, et pluseours autres rebelz des voz marches de Galys ... **Warfore, for goddesake, thinketh on your beste frende, god, and thanke hym as he hath deserved to yowe! And leueth nought that ye ne come for no man that may counsaile yowe the contrarie** ... Tresexcellent, trespuissant, et tresredouté Seignour, autrement say a present nieez. Jeo prie a la benoit trinité que vous otroic bone vie ovc tresentier sauntee a treslonge durré, and **sende yowe sone to ows in help and prosperitee; for in god fey, I hope to almighty god that, yef ye come youre owne persone, ye schulle have the victorie of alle youre enemies** ... Escript a Hereford, en tresgraunte haste, a trios de la clocke apres noone. le tierce jour de Septembre.*

'May your most gracious Lordship be **pleased** to hear that today, in the afternoon ... more than 400 of Owen, Glyn and Talyard's rebels, and several other rebels from your Welsh borders, entered our **county**. **Wherefore, for God's sake, set your mind on God as your best friend, and thank him for the favours he has bestowed upon you. And do not for any reason fail to come, whatever advice to the contrary you may receive from anyone** ... Most excellent, most powerful and most redoubtable Lord, let me be denied / refused in some other way! I pray the blessed Trinity that you be granted good life with perfect health for a long time to come, **and may [the Trinity] send you to us soon in help and prosperity; for I faithfully pray to almighty God that, if you yourself come in person, you will be victorious over all your enemies** ... Written [by R. Kingston to King Henry IV] at Hereford in the utmost haste at three o'clock in the afternoon on the third day of September [1403].'

Schendl [2002] argues that the frequent use of code-switching in documents such as letters to the king, is an indication of the social acceptability of code-switching in Medieval times.

Swedish examples of code-switching are found in the Gender and Work corpus of court records and church documents from the Early Modern period (see further Chapter 3.1 for more details on the contents of the corpus). The following example of code-switching between Swedish and Latin is from a

church ordinance from 1571 (Latin in **bold**):

*När tijdh är gå **Ordinati ad sacram Comunionem***

‘When time is go **Ordinati ad sacram Comunionem**’

‘When it is time, **the ordained go to the Holy Communion**’

Code-switching also occurs in the court records part of the same corpus, as in the following example from *Revsunds tingslag* (1649–1689) (Latin in **bold**):

*...begärandes wetta **quo iure** dee dem innehafwa*

‘...requiring to know **quo iure** they them have’

‘...requiring to know **by what right** they are in possession of them’

To sum up, NLP for historical text is a tricky task. Nevertheless, there are ways of dealing with the special characteristics of historical source material, enabling linguistic analysis and further NLP processing of historical text. In the remaining part of this chapter, I will discuss some of the work conducted within this area, and also relate these findings to similar research performed for modern language data.

2.2 Adapting NLP Tools to the Historical Domain

A straightforward way of enabling linguistic analysis of historical text is to perform domain adaptation by training taggers and parsers on annotated historical data, resulting in NLP tools adapted to the specific domain of a historical language variant. Sánchez-Marco et al. [2011] tried a strategy for adapting the existing FreeLing tool to the task of morphosyntactic tagging of Old Spanish (12th to 16th century). In its original architecture, the FreeLing tool is designed to perform NLP tasks such as tokenisation, tagging, parsing, and semantic analysis for contemporary standard Asturian, Catalan, English, Galician, Italian, Portuguese, Russian, Spanish, and Welsh [Padró and Stanilovsky, 2012]. As training data in adapting the tagging module in FreeLing to the Old Spanish domain, a corpus of approximately 20 million tokens was used, containing texts from the 12th to the 16th century, distributed over eight different fiction and non-fiction genres. Based on this corpus, an Old Spanish tagging module for the FreeLing tool was developed by:

1. extending the original (contemporary) Spanish dictionary with Old Spanish word forms extracted from the corpus and mapped to their corresponding modern spelling,

2. extending the original set of affixation rules for handling unknown word forms with Old Spanish derivational affixes,
3. modifying the tokenisation module to handle specific characteristics of the Old Spanish corpus, and
4. retraining the tagger on a gold standard subset of the Old Spanish corpus, containing 30,000 tokens pre-annotated with the modern Spanish tagger and then manually revised and corrected.

The evaluation results showed that the dictionary expansion part of the adaptation had the largest positive impact on the results. In the best-performing setting, the tagger achieved an accuracy of 94.5% in finding the right part of speech for Old Spanish input data, and 89.9% in finding the complete morphological tag.

Rögnavaldsson and Helgadóttir [2011] implemented a bootstrapping technique for adapting a tagger originally developed for Modern Icelandic to the task of morphosyntactic tagging of Old Icelandic text. The tagger used in the experiments is the TnT tagger [Brants, 2000], trained on the IFD corpus of approximately 590,000 tokens from the 1980s. The corpus used for adaptation to historical text, and evaluation of the results, consists of narrative prose texts (sagas), presumably written in the 13th and 14th centuries. From this corpus, containing in total 1,650,000 tokens, a training set of 95,000 tokens was randomly sampled, and an evaluation set of 4,000 tokens. The adaptation of the TnT tagger to Old Icelandic input text was conducted in the following steps:

1. **Generation of Baseline Results**

The original tagger developed for Modern Icelandic was applied to the Old Icelandic evaluation corpus, yielding a baseline accuracy of 88.0% correct tags.

2. **Semi-automatic Annotation of the Training Corpus**

In the second step, the Old Icelandic training corpus was pre-tagged using the Modern Icelandic tagger. The resulting morphosyntactically annotated text was then manually revised and corrected.

3. **Re-training**

In the third step, the tagger was re-trained on the 95,000 manually revised and corrected Old Icelandic tokens. This resulted in a tagging accuracy of 91.7%, when applying the re-trained tagger to the evaluation corpus.

4. **Combining Old and Modern Data**

Finally, the tagger was trained on the union of the manually corrected Old Icelandic training corpus and the Modern Icelandic IFD corpus. This approach was inspired by related work on developing NLP tools for less-resourced languages by training taggers and parsers for one lan-

guage on data from a closely related language. Using this combined model, trained on both Old Icelandic data and Modern Icelandic data, a tagging accuracy of 92.7% was reported for the Old Icelandic evaluation corpus.

An overall aim of their work on developing a tagger for Old Icelandic, was to facilitate linguistic studies on syntactic variation and change. To see whether their tagger would be suitable for this task, they used their automatically tagged version of the Old Icelandic corpus to search for two disputed features of Old Icelandic syntax: *object shift* and *derivational passive*.

The object shift allows for a direct or indirect object to move to the left of a negation, as in the following two examples given by Rögnvaldsson and Helgadóttir [2011]:

Nemandinn las *bókina ekki*
the-student read *book not*
'The student didn't read the book'

Nemandinn las *hana ekki*
the-student read *she not*
'The student didn't read it'

In Modern Icelandic, object shifts may occur for pronouns as well as for full noun phrases, whereas for the closely related languages Danish, Norwegian, and Swedish, this process only applies to pronouns. Researchers believe that this property of the Icelandic language is related to its rich case morphology as compared to the other Scandinavian languages. For this claim to be true, object shifts applied to full noun phrases would be expected to occur in Old Icelandic text as well, since the case morphology of Modern Icelandic is more or less equal to the case morphology of Old Icelandic. It has however been stated that this kind of object shift does not occur in Old Icelandic text, which would contradict this theory. Rögnvaldsson and Helgadóttir [2011] searched their morphosyntactically tagged Old Icelandic corpus for this kind of syntactic construction, and ended up with 9 unquestionable examples of noun phrase-based object shift. Likewise, they used the same corpus to provide evidence for the existence of derivational passive in Old Icelandic, another previously disputed phenomenon.

For parsing, Schneider [2012] developed a method for adaptation of a contemporary English parser to the domain of historical English. The parser used in the experiments is the Pro3Gres dependency parser, based on hand-written grammar rules combined with statistical disambiguation models derived from the Penn Treebank [Schneider, 2008]. The texts used for evaluation were obtained from the Archer corpus (A Representative Corpus of Historical English Registers), containing British and American English text from the time pe-

riod 1600–1999 [Biber et al., 1994]. Adaptation of the parser to the historical English domain was performed in four steps:

1. **Spelling Normalisation**

The spelling in the historical input text was automatically converted to a more modern spelling, using the VARD normalisation tool (see further Section 2.3.2) with a pre-existing dictionary covering Early Modern English spelling [Baron and Rayson, 2008].

2. **Grammar Extension**

The hand-written grammar rules were extended to cover unexpected interpretations of specific word forms.

3. **Punctuation Removal**

All the commas were removed from the input text, in order not to confuse the parser due to a different, and inconsistent, use of punctuation in the historical data.

4. **Relaxed Word Order Constraints**

The word order constraints defined in the original parser were relaxed with the aim of better coping with the freer word order in historical text. More specifically, the constraints on fronted prepositional phrases were relaxed, to allow for this type of construction in all contexts (not only at sentence-initial position).

Evaluation was performed on a small subset of the ARCHER corpus, containing in total 100 sentences distributed over 25 sentences each from the 17th century, the 18th century, the 19th century, and the 20th century. Evaluation scores were given in precision and recall for the parsing relations *subject*, *object*, *PP-attachment*, and *subordinate clauses*. From the experiments, it was concluded that spelling normalisation was the best-performing adaptation in this setting. The grammar extension adaptation on the other hand, was relatively easy to implement, due to the grammar being hand-written, but led to small improvements only. The more global adaptations, in the form of removing commas and relaxing word order constraints, were successful in handling some problematic cases, but also introduced new parsing errors. Including all four adaptations, precision and recall scores of approximately 70% were achieved for the 17th century part of the corpus, as compared to approximately 80% for the 20th century part.

2.3 Applying Modern NLP Tools to Historical Data

To be able to train taggers and parsers specifically aimed at the task of analysing historical text, linguistically annotated historical corpora are required. There are some reasonably large corpora of this kind available, such as the Penn

Parsed Corpora of Historical English [Kroch and Taylor, 2000, Kroch et al., 2004], the Icelandic Parsed Historical Corpus [Rögnvaldsson et al., 2012], the Tycho Brahe Parsed Corpus of Historical Portuguese [Galves and Faria, 2010], the Gold Standard Corpus of Early Modern German [Scheible et al., 2011a], and others. However, for most languages, such corpora are non-existing or very limited in size. Due to this lack of annotated historical data, combined with the large orthographic and syntactic variance in historical text, training taggers and parsers for the historical domain is often not an option. To still be able to perform linguistic analysis of historical text, there have been experiments on applying NLP tools available for the modern language to the task of analysing historical text, with or without pre-processing of the input data. Pre-processing of the input text is most commonly performed with the aim of modernising the spelling in the historical input data before applying the modern NLP tools, a process often referred to as *spelling normalisation*. The different methods for spelling normalisation, some of which are presented in the following sections, have to a great extent been inspired by methods traditionally used in fields such as spelling correction, speech technology, and machine translation. It could also be noted that modern text, such as Twitter and SMS text, display similar spelling variance as historical text. Thus, similar normalisation techniques have also been applied to these text types, which is further discussed in Section 2.4.

2.3.1 Applying Tools Off-the-Shelf

Even though historical text differs from modern text in many respects, there are still similarities, and there have been attempts to apply contemporary taggers and parsers directly to the historical input text, without any adaptation neither of the input text nor of the tool in itself. Pennacchiotti and Zanzotto [2008] evaluated dictionary coverage and part-of-speech tagging results for contemporary Italian resources when applied to a corpus of Italian text from the time period 1200–1881. For the dictionary coverage experiments, two resources were included:

1. A manually built generative morphology dictionary, covering approximately 22,000 lemmas corresponding to a total of nearly 74,000 different word forms.
2. A corpus-induced dictionary extracted from a collection of articles from an Italian financial newspaper, containing in total around 12,000 word forms.

The results showed that only 27,3% of the word forms in the oldest text (from the year 1200) were covered in the dictionary, as compared to 48,7% in the youngest text sample (from the year 1881). Comparisons were also made to a modern language corpus extracted from the Italian newspaper *La Repubblica*,

in which approximately 62,5% of the word forms were covered by the dictionary, and a large proportion of the unknown word forms were proper names.

For the part-of-speech tagging experiments, the Chaos tool developed for Modern Italian was used [Basili and Zanzotto, 2002]. For evaluation of part-of-speech tagging results, a manually annotated gold standard set of 42 sentences was compiled from the historical corpus. When applying the contemporary Chaos tagger to these 42 sentences, a tagging accuracy of 54% was achieved for the oldest text, as compared to 68% for the youngest historical text, and 97% for the modern corpus extracted from *La Repubblica*. The authors concluded that NLP tools developed for the modern language are not suitable for being applied directly to historical text. Several ways of improving the performance were suggested, such as adding historical dictionaries to the analysis process, or using Levenshtein comparisons to map historical word forms to modern word forms.

Similar conclusions were drawn by Scheible et al. [2011b] when evaluating an off-the-shelf part-of-speech tagger developed for Modern German on Early Modern German text. In their experiments, they applied the TreeTagger [Schmid, 1994] to a manually annotated subset of the GerManC corpus [Scheible et al., 2011a], containing nearly 58,000 tokens from the time period 1650–1800. Two versions of the corpus were created; one with the original spelling preserved and one where the spelling had been manually normalised into a modern spelling. When applying the TreeTagger to the unnormalised version of the corpus, a tagging accuracy of 69.6% was achieved, as compared to 79.7% for the modernised version of the text, which is still far from the approximately 97% previously reported for the TreeTagger on modern German text.

2.3.2 Dictionary-based Adaptation

Since the performance of modern NLP tools on historical text has proven to be low (as argued in the previous section), and there is a general lack of annotated historical data available for training NLP tools adapted to the historical domain, other solutions are called for. One such solution is to include dictionaries in a pre-processing step, with the aim of mapping historical word forms to their modern spelling, thus making it easier for the NLP tools to relate the input form to previously seen (modern) training data.

An early example of adding historical dictionaries to boost the performance of modern NLP tools when applied to historical text was presented by Rocio et al. [1999]. The aim of their work was to perform syntactic annotation of Medieval Portuguese (13th to 14th century). Due to the lack of Medieval Portuguese language resources, they applied a phrase structure chart parser grammar developed for contemporary Portuguese to the task of syntactically annotating Medieval Portuguese text. The parser contained 250 grammar rules

and a dictionary of approximately 1,000,000 contemporary Portuguese entries [Rocio and Lopes, 1999]. To deal with differences in spelling between Medieval Portuguese and contemporary Portuguese, a small dictionary was added, containing articles, pronouns, certain verbs, and a set of inflectional rules. The method was used as a pre-annotation step for manual syntactic annotation of Medieval Portuguese. The authors concluded that their approach was successful for the task of partial parsing of medieval Portuguese texts, even though there were some problems remaining concerning grammar limitations, dictionary incompleteness and insufficient part-of-speech tagging.

One of the most widely used dictionary-based approaches to spelling normalisation of historical text is the VARIant Detector (VARD), developed by Rayson et al. [2005]. The VARD tool was originally implemented for Early Modern English (~1450–1700), and consists of a mapping scheme from old spellings of a word form to the corresponding modern spelling. Whenever a text is run through the system, all word forms that are found in the dictionary are replaced by their modern spelling variant, facilitating further linguistic processing using modern NLP tools. The VARD dictionary was created on the basis of three Early Modern English sources:

1. Newspaper text from 1653–1654.
2. *The Nameless Shakespeare* – a collection of Shakespeare plays available in different spellings.
3. Texts extracted from *the Eighteenth and Nineteenth Century Fiction collection*.

From this text material, a list of varying spelling forms was generated by manual inspection of the texts, with special focus on words flagged as unknown by a tagger. In addition, *The Oxford English Dictionary* and other historical sources were consulted for extending the list of possible spelling variants for a particular word form. In total, the resulting dictionary comprises 45,805 entries, mapping historical word forms to their modern spelling.

The VARD tool was evaluated by comparing its ability to correctly identify and replace old spellings by modern spellings in Early Modern English text to the performance of two modern spell checkers for the same task. The spell checkers used for this study were Microsoft Word 2002 and Aspell. The evaluation was performed on a collection of texts from the time period 1666–1679, extracted from the *Lampeter Corpus of English Tracts* [Schmied, 1994].

The results showed that between a third and a half of all tokens (depending on which test text was used) were correctly normalised by both VARD and Microsoft Word, whereas approximately 29.4% of the tokens were correctly normalised only when using VARD. The percentage of tokens correctly normalised only by Microsoft Word was substantially lower; approximately 6.7%. The comparison between VARD and Aspell showed similar results.

An extrinsic evaluation was also performed, evaluating the performance of the CLAWS tagger on Early Modern English text from Shakespeare and from the Lampeter corpus. This extrinsic evaluation showed that part-of-speech tagging accuracy improved from 81.9% to 84.1% for the Shakespeare text, and from 88.5% to 89.4% for the Lampeter corpus, when using VARD for automatically modernising the spelling before the tagger was applied [Rayson, 2007].

VARD was later further developed into VARD2, combining the original word list approach with techniques derived from modern spell checking programs [Baron and Rayson, 2008]. The full VARD2 system thus includes three modules:

1. The manually defined mapping scheme between old word forms and their modern spelling.
2. A phonetic matching module based on SoundEx, mapping old word forms to their modern spelling based on phonetic similarity comparisons.
3. A set of letter replacement rules for handling frequently occurring spelling variation in Early Modern English text, such as the doubling of certain characters or replacing the letter *v* by the letter *u*.

In Section 3.4, I present and evaluate my version of the dictionary-based approach to spelling normalisation, which I refer to as *memory-based normalisation*, since the dictionary used in my approach is more similar to a translation memory than to a traditional dictionary.

2.3.3 Levenshtein-based Spelling Normalisation

Even though dictionary lookup yields promising results in the context of applying NLP tools developed for the modern language to historical text, this method is highly dependent on the size and coverage of the dictionary. Another technique traditionally implemented for spell checking, and also applied to the task of spelling normalisation of historical text, is the introduction of Levenshtein edit distance comparisons. The Levenshtein edit distance is used to calculate the similarity between two strings by counting the number of characters that need to be deleted, inserted, or substituted in order to transform one string into the other [Levenshtein, 1966].

One Levenshtein-based approach to spelling normalisation of historical text was developed for Early New High German (14th to 16th century) by Bollmann et al. [2011]. In this approach, normalisation rules are automatically learned from word-aligned parallel corpora of historical and modern text. In their experiments, they used two versions of the Martin Luther bible: the Early New High German version from 1545, and a New High German version from

1892 with a modernised spelling and with extinct word forms replaced by modern word forms.

To extract word-to-word mappings from this parallel bible corpus, the sentences and the words in the older corpus need to be linked to the corresponding sentences and words in the newer corpus. For sentence alignment, the Gargantua toolkit was used [Braune and Fraser, 2010], whereas word alignment was performed using the GIZA++ toolkit [Och and Ney, 2003]. After removing alignments that were suspected to be incorrect, for example due to large differences in string length between the source string and the target string, a corpus of approximately 550,000 aligned word pairs was produced. From this corpus, 40% of the alignment pairs were randomly selected for a training corpus, used for extracting replacement rules based on the edit distances between the strings in the Early New High German version of the bible as compared to their spelling in the modern version of the bible. Likewise, another 20% of the alignment pairs were randomly selected for a development corpus, whereas yet another 20% were used for evaluation.

The set of normalisation rules induced from the training corpus were ranked according to their frequency when applied to the development part of the corpus. This rank was then used as a probability score in the normalisation process. In addition, a modern language dictionary was used to avoid the generation of non-existing word forms. Thus, from the list of normalisation candidates generated by the edit distances comparison rules, only those word forms that were also found in the modern language dictionary were kept for further processing. From the remaining normalisation candidates, the one with the highest probability score was chosen as the final normalisation candidate.

This Levenshtein-based approach to spelling normalisation of Early New High German was evaluated in terms of ‘identical tokens’, that is the number of word forms that are identical in the automatically normalised version of the text as compared to the modern version of the bible. Their results showed that approximately 64,7% of the word forms in the historical, unnormalised version of the text already had a spelling that was identical to the modern gold standard spelling. Applying the Levenshtein-based normalisation rules combined with a dictionary for avoiding normalisation into non-existing word forms, resulted in an increase to 91.0% word forms with a spelling identical to the modern version.

Later, Bollmann [2012] tried a combination of dictionary lookup and different modifications to the original Levenshtein distance measure, improving normalisation accuracy further to 92.6% for the same training and test corpora. An extrinsic evaluation was also performed, comparing the performance of the RFTagger applied to historical text before and after normalisation. For every evaluation text, the tagger was trained on between 100 and 1,000 manually normalised tokens, and evaluated on the remaining tokens in the same text. For one manuscript from the 15th century, tagging accuracy was improved

from 28.7% to 78.0% using this method, which could be compared to 87.1% for the perfectly normalised version of the text [Bollmann, 2013].

For the Swedish language, Borin et al. [2007] proposed the use of an existing, contemporary named-entity recognition system for analysing Swedish literature from the 19th century. The system was designed to recognise person names, locations, organisations, artifacts (food and wine products, vehicles etc), Work&Art (names of novels, sculptures etc), events (religious, cultural etc), measure/numerical expressions and temporal expressions. In the first run, the named entity recognition system was applied to texts from the Swedish Literature Bank without any adaptation to historical input data, resulting in problems with spelling variation, inflectional differences, unknown names and structural issues (such as hyphens splitting a single name into several entities). In the basic version of the system, relying solely on dictionary lookup, the application to Swedish 19th century texts, without any domain adaptation, yielded an F-score of 78.1% for the task of named entity recognition. In the second run, they tried an approach where the text was normalised to a modern spelling, before the named entity recognition system was applied. The spelling normalisation step was then performed by adding alternate spelling variants to the dictionary, and by including edit distance comparisons between possible name mentions and the contemporary dictionary. The inclusion of the spelling normalisation step resulted in an increase in F-score to 89.5% for the same task.

In Section 3.3, I present and evaluate my version of the Levenshtein-based approach to spelling normalisation, where edit distance comparisons are made between the original historical word form and entries in a modern language dictionary.

2.3.4 Spelling Normalisation based on Phonetic Similarity

Another similarity-based method for spelling normalisation of historical text was suggested by Jurish [2008], and referred to as *conflation by phonetic form*. He argued that due to the lack of orthographic conventions in historical time, spelling generally reflects the phonetic form of the word to a higher degree in historical text. Furthermore, it is assumed that phonetic properties are less resistant to diachronic change than orthography. Accordingly, he explored the idea of comparing the similarity between phonetic forms in the normalisation process, instead of comparing orthographic forms. For mapping graphemes in the written language to phonemes in the spoken language, he used an existing rule set for letter-to-sound conversion for Modern German, distributed with the IMS German Festival text-to-speech system [Black and Taylor, 1997]. This rule set was adapted to historical German by including rules such as ignoring the grapheme *h* (assuming it to be silent) and mapping successive occurrences of any vowel to a single occurrence of the same vowel.

Evaluation of the phonetically based approach to spelling normalisation was performed on a corpus of historical German verse quotations extracted from *Deutsches Wörterbuch*, containing approximately 5,500,000 tokens, corresponding to $\sim 320,00$ types. Without normalisation, approximately 83.7% of the tokens were recognised by a morphological analyser. After normalisation, 91.6% of the tokens were recognised. Adding lemma-based heuristics, coverage increased further to 94.4%.

In my experiments, I have not yet tried the phonetic similarity approach to spelling normalisation. I do however find this method interesting for future work.

2.3.5 SMT-based Spelling Normalisation

Another way of viewing the issue of spelling variation in historical text, is to treat spelling normalisation as a translation problem. Scherrer and Erjavec [2013] tried character-based statistical machine translation (SMT) techniques for modernising the spelling in Slovene texts from the 18th and 19th century. In the SMT-based approach to spelling normalisation, probabilistic models are used for finding the most probable string in the target language corresponding to a string observed in the source language. In this context, the probabilistic model consists of two fundamental components: the *translation model* and the *language model*. The translation model is used for estimating the likelihood that certain units in the target language are translations of a string in the source language. For this purpose, a word-aligned parallel corpus of the source and target language is needed. The language model, on the other hand, estimates the likelihood that a certain string would occur in the target language, thus contributing to the fluency and idiomaticity of the translation. For this purpose, a monolingual corpus of the target language is used [Koehn, 2010].

In the character-based setting, the models are trained for translating character sequences rather than words and phrases, a method that has previously been successfully implemented for transliteration and translation between closely related languages [Matthews, 2007, Vilar et al., 2007, Tiedemann and Nabende, 2009, Tiedemann, 2009, Nakov and Tiedemann, 2012]. Scherrer and Erjavec [2013] applied the character-based SMT technique to the task of spelling normalisation of historical Slovene in two different experiments:

1. Supervised learning

In the supervised setting, the translation model was trained on a set of 45,810 historical-modern Slovene word pairs, whereas the language model was trained on the same data set but only including the modern word forms. In addition, a lexicon filter was used, in which normalisation candidates proposed by the translation model were only accepted if they were also found in the Modern Slovene Sloleks dictionary (containing 930,000 entries).

2. Unsupervised learning

In the unsupervised setting, no historical-modern training data is presupposed. Instead this kind of training data is created in an unsupervised manner, based on separate lists of historical word forms and modern word forms. The historical word forms are mapped to modern word forms based on string similarity comparisons between the words occurring in the two lists, using the BI-SIM measure [Kondrak and Dorr, 2004]. In all other respects, the setup was equal to the supervised setting, with the same data for language modeling, and with the inclusion of a lexicon filter.

Evaluation results were presented in terms of normalisation accuracy, based on an evaluation corpus of historical word forms mapped to their manually modernised spelling, and calculated separately for three time periods: the second half of the 18th century, the first half of the 19th century, and the second half of the 19th century. As a baseline, they used the number of word forms in the original input text that already have a spelling identical to the manually modernised spelling. As could be expected, the amount of words with a modern spelling varies between the different time periods. In the 18th century corpus, only 15.4% of the word forms have a modern spelling, as compared to 82.5% of the word forms in the late 19th century corpus. Accordingly, spelling normalisation had the largest effect on the oldest subcorpus, increasing accuracy from 15.4% to 48.9% in the unsupervised setting, and to 72.4% in the supervised setting. For the youngest subcorpus, normalisation accuracy increased from 82.5% for the original text to 87.4% in the unsupervised setting, and to 92.7% in the supervised setting.

In Section 3.5, I present and evaluate my version of the character-based SMT approach to spelling normalisation.

2.4 Relation to Spelling Variation in Modern Data

Throughout this chapter, I have given an overview of common approaches to handling spelling variation in the context of NLP for historical text, with a few representative examples of research given for each approach presented. In this context it should also be noted that the characteristics of historical text have many similarities with the characteristics of modern text such as SMS, chat and Twitter text. This includes a high degree of spelling variation, ungrammatical sentences and ad hoc elisions and abbreviations. As for historical text, a common approach to enable the application of NLP tools to SMS and Twitter data is thus to perform spelling normalisation, transforming the input data to canonical word forms consistent with the standard language, before applying taggers and parsers to the text. As a consequence, many of the approaches

to spelling normalisation of historical text presented in this chapter have also been implemented in a similar manner for SMS and/or Twitter data.

Bilal [2014] used a combination of Levenshtein edit distance calculations and refined Soundex matching for normalising Twitter data. In his approach, all word forms were first matched against a standard language dictionary. For all word forms not found in the dictionary, Levenshtein edit distance comparisons between the original word form and the word forms present in the dictionary were used for generating a list of possible normalisation candidates. In a second step, the refined Soundex algorithm was used to rank the list of candidates based on phonetic similarity.

Pennell and Liu [2011] developed a method for spelling normalisation of SMS messages using character-based SMT techniques. As a training corpus for the translation model, a set of 4,661 Twitter status messages were manually normalised. In addition, a language model was used for choosing a hypothesis based on the context. They showed a normalisation accuracy of 60.4% for the top-1 suggestions given by the system, and 75.6% when taking the top-20 suggestions into consideration.

Han and Baldwin [2011] presented a method for normalising SMS and Twitter text based on morphophonemic similarity, combining features such as lexical edit distance, phonemic edit distance, prefix substring, suffix substring, and the longest common subsequence. Context was taken into account by means of dependency structures generated by the Stanford Parser applied to a corpus of New York Times articles. In the best setting, a token-level F-score of 75.5% and 75.3% was reported for SMS messages and Twitter texts respectively.

These were just a few representative examples showing that the same kinds of techniques have been used for enabling the application of NLP tools to historical data and to modern non-standard data. In both cases, the main idea is to adapt the input data to the NLP tools, rather than adapting the NLP tools to varying and inconsistent input data. In the second part of this thesis, I will present and evaluate my four approaches to spelling normalisation of historical text, partly inspired by the dictionary-based, the Levenshtein-based, and the SMT-based spelling normalisation techniques discussed throughout this chapter.

Part I:
Spelling Normalisation of Historical Text

3. Approaches to Spelling Normalisation

Spelling normalisation is a crucial step in my approach to linguistic analysis and information extraction from historical text. By spelling normalisation I refer to the process of automatically converting historical word forms to a modern spelling, enabling the application of modern NLP tools to the text in the next step. For illustration, consider the following Middle English segment:

To the moost noble & Worthiest Lordes moost ryghtful & wysest conseilie

In the *Innsbruck Computer Archive of Machine-Readable English Texts* [Markus, 1999], this sentence is normalised as:

To the most noble and Worthiest Lords most rightful and wisest council

In the original segment, the word forms *moost*, *ryghtful*, *wysest* and *conseilie* would probably be unknown to a tagger or parser developed for Modern English, and thus potentially cause an incorrect linguistic analysis of the segment. After spelling normalisation, a higher proportion of the word forms are likely to be recognised by the NLP tools, with a better chance of arriving at a correct linguistic analysis, even though there may still be structural differences as compared to a sentence originally written in Modern English.

Depending on the available resources for a certain language and time period, different normalisation methods could be used. I try four approaches to spelling normalisation, with different requirements concerning the level of human expertise and language resources needed for implementation:

1. rule-based normalisation
2. Levenshtein-based normalisation
3. memory-based normalisation
4. SMT-based normalisation

The rule-based approach was developed as a proof-of-concept, to test the feasibility of spelling normalisation in this context. In the rule-based approach, rewrite rules are manually defined for transforming historical spelling into modern spelling, based on known spelling characteristics in text from a certain period of time. Even though this method is language-dependent and specifically aiming at normalising text from a particular time period, it has the advantage that no annotated historical training data is needed, provided that there is some knowledge about what spelling differences the rules should cover.

In contrast to the rule-based approach, the Levenshtein-based approach to spelling normalisation does not rely on human expertise for creating normalisation rules. Instead, normalisation is performed based on string similarity comparisons between the historical word form to be normalised and word forms occurring in a modern language dictionary. The main advantages of the Levenshtein-based approach are that it is applicable to any language for which a modern language dictionary exists, and that no human expertise on spelling characteristics for a specific language or time period is needed for implementation. On the other hand, the Levenshtein-based approach is highly dependent on the coverage of the modern language dictionary used for string similarity comparisons, as word forms that are not part of the dictionary will be out of reach for this normalisation approach. This is problematic, since the variable nature of language means that the full vocabulary of a language can never be covered in a dictionary. Furthermore, historical language may contain words that are no longer part of a modern vocabulary, such as archaic word forms and occupational titles no longer in use.

Both the rule-based approach and the Levenshtein-based approach are applicable in cases where no historical data is available for training. If historical text is accessible both in its original spelling and in a manually validated modern spelling, data-driven normalisation methods may also be applied. In the memory-based approach, normalisation is performed by comparing each historical word form to word forms occurring in a parallel training corpus of historical word forms mapped to their manually normalised counterparts. In the normalisation process, word forms that are found in the training corpus are substituted by the most frequent normalisation form occurring in the corpus, whereas word forms not found in the corpus are left unchanged. The main advantage of this approach is its simplicity. Provided that there is a historical-modern parallel corpus available, no language-specific knowledge is needed for implementing the normalisation process. The memory-based method is also generic in the sense that it is applicable to any language for which such a corpus exists. One obvious disadvantage is that this kind of corpus is not available for all languages and time periods. Similar to the Levenshtein-based approach, the memory-based approach is furthermore highly dependent on the coverage of the language resource used for training, since any word form that is not present in the parallel training corpus will be out of reach for this normalisation method.

Parallel corpora are otherwise often associated with the field of machine translation, where statistical machine translation (SMT) systems are trained on parallel corpora of manually translated documents. In the SMT-based normalisation approach, I apply standard SMT techniques for training a historical-to-modern machine translation system. As training data I use the same kind of corpus as in the memory-based approach, that is a parallel corpus of historical word forms mapped to their manually modernised spellings. Since I want to handle spelling variation rather than the translation of words and

phrases, the system is trained on character sequences instead of word sequences. Advantages of the SMT method are that the SMT techniques are language-independent, and no human knowledge about spelling differences is required for implementation. Furthermore, since the SMT-based approach operates at a character level, previously unseen word forms may also be generated in the normalisation process. A possible disadvantage of the SMT-based method is that in traditional machine translation setups, large parallel training corpora are required to achieve high quality translations. Since the spelling normalisation system is trained on the much more limited set of character sequences instead of word sequences, less training data is however needed for this task than for conventional machine translation tasks.

In the following, I start by presenting the Swedish Gender and Work corpus, which is the main source used for training and evaluation of the four approaches to spelling normalisation presented above. Thereafter, each normalisation method is described in more detail, and evaluated. The aim of the evaluation is to explore optimal settings for each approach and/or how well the methods comply with the criteria of being generally applicable to different time periods and genres, as well as not requiring large amounts of annotated historical data for training. In addition, Chapter 4 presents a joint evaluation of the three language-independent approaches when applied to historical data from five different languages. The aim is to explore the applicability of the normalisation methods to other languages, and to investigate whether the approach that yields the best results for Swedish is the best-performing setup for other languages as well. A summary and conclusions from the normalisation experiments are given in Chapter 5.

3.1 The Gender and Work Corpus

For training and evaluation of the different approaches to spelling normalisation, I mainly use Swedish data available within the Gender and Work project (further described in Chapter 1). The subset of the Gender and Work material used in my experiments, henceforth referred to as the *Gender and Work corpus*, is a collection of 11 court records texts and 4 church documents from the time period 1527–1812, comprising a total of 787,122 tokens. From this corpus, 40 randomly selected sentences from each text (in total 600 sentences) were extracted to a training set, and another equally sampled 600 sentences to an evaluation set. The proportion of tokens in each text as a whole, as well as in the training and evaluation parts, is given in Table 3.1.

To enable training and evaluation of the different normalisation techniques, I manually normalised all the tokens in the training and evaluation parts of the corpus, mapping each historical word form to its modern language equivalent. In most cases, this manual normalisation process is a rather intuitive and straightforward task, as when mapping the historical spellings *aff*, *wara* and *til*

Court Records				
Name	Time Period	Total	Training	Evaluation
Östra Härads i Njudung	1602–1605	38,477	1,980	2,069
Vendels dombok	1615–1645	64,977	1,583	2,509
Per Larssons dombok	1638	12,864	2,848	2,987
Hammerdals tingslag	1649–1686	75,143	1,508	1,859
Revsunds tingslag	1649–1689	113,395	2,275	2,328
Stora Malm	1728–1741	458,548	1,627	1,895
Vendels dombok	1736–1737	61,664	3,032	3,450
Stora Malm	1742–1760	74,487	2,034	2,336
Stora Malm	1761–1783	66,236	1,905	1,825
Stora Malm	1784–1795	58,738	2,036	1,378
Stora Malm	1796–1812	47,671	2,345	1,683
Church Documents				
Name	Time Period	Total	Training	Evaluation
Westerås recess	1527	14,149	2,831	3,709
Swenska kyrkoordningen	1571	60,354	2,093	2,246
Uppsala Möte	1593	34,877	1,070	1,184
Kyrkolag	1686	35,201	1,660	2,086
Total	1527–1812	787,122	30,827	33,544

Table 3.1. *Corpus distribution for the Gender and Work corpus. Total = Number of tokens in the text as a whole. Training = Number of tokens extracted to the training set. Evaluation = Number of tokens extracted to the evaluation set.*

to their modern spellings *av* ('of/off'), *vara* ('to be') and *till* ('to'). The main guideline in the normalisation process is that the normalised version should be a word form that is likely to be present in a modern language dictionary. Exceptions are archaic word forms, such as the previously discussed occupational title *brofogde* ('surveyor of bridges'), which would normally not occur in a modern language dictionary since this occupation does not exist in the modern society. In such cases, it is however often clear to a native speaker of Swedish how this word form should be spelled, and I therefore normalise these word forms according to intuition. For example, in the Gender and Work corpus, this particular occupational title is spelled as *brofougde*, whereas I suggest the modern spelling *brofogde*. In cases where it is unclear to me what the modern spelling of a word form would be, I leave the word form in question unchanged.

Regarding proper nouns, I have decided to normalise names of places, but not person names. The motivation for this is that names of places are standardised to a larger extent than person names, and the same places exist today, with a modern spelling variant. This is why I normalise for example the city name *Upsala* into its modern spelling *Uppsala*, whereas person names such

as *Oluff* and *Swen* are left unchanged, although these would typically (but not always) be spelled *Olof* and *Sven* in contemporary Swedish.

Even though I refer to this process as *spelling normalisation*, I sometimes include more than spelling in the mapping between old and modern word forms, to cover aspects such as morphological change. One example is the historical word form *närvarellse*. If only spelling was taken into account, this word form would typically be normalised into *närvarelse*. This is however not a word that is part of the modern vocabulary. Due to morphological change, the modern equivalent would instead be *närvaro*, which hence is the word form chosen for normalisation. Likewise, I have chosen to normalise the historical word form *tilbragte* into the modern word form *tillbringade*, which involves a change in inflectional paradigm in addition to a difference in spelling.

3.2 Rule-based Normalisation

In the rule-based approach to normalisation, spelling transformation rules are manually defined on the basis of known or observed spelling changes in the history of a specific language. In my experiments, I define a set of 29 hand-crafted rules, based partly on the reformed Swedish spelling introduced in 1906 [Bergman, 1995], and partly on the initial 20 sentences (984 tokens) in *Per Larssons dombok*, a court records text from 1638 [Edling, 1937], which is also part of the Gender and Work corpus. The rule-based setup is illustrated in Figure 3.1.

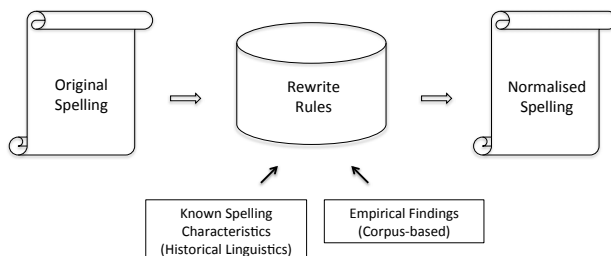


Figure 3.1. Overview of the rule-based spelling normalisation workflow.

Rules based on the reformed spelling include for example:

- the simplification of the *t* sound from *dt* to a single *t*:
varidt → *varit* ('been')
- the drop of the superfluous letters *h* or *f* for the *v* sound:
hvar → *var* ('was')
skrifva → *skriva* ('write')

- the abandonment of using an *f* for denoting a *v* sound:
af → *av* ('of/off')

Examples of rules based on the court records text are:

- the substitution of letters to a phonologically similar variant:
qvarn → *kvarn* ('mill')
slogz → *slogs* ('were fighting')
- the deletion of repeated vowels:
saak → *sak* ('thing')
- the deletion of mute letters:
vijka → *vika* ('fold')
- the Swedification of spelling influenced by other languages (mainly German):
schall → *skall* ('shall')

The full set of rules is given in Appendix A.

Due to the biased text material used as a basis for developing the normalisation rules, it could be expected that the rule-based approach would yield better normalisation results for 17th century text than for texts from other time periods. As there were no standardised spelling conventions during this period, another hypothesis is that there will be differences in the applicability of the rules between texts written by different authors and within different text genres. In the following, I explore the impact of the rule-based approach on the Gender and Work corpus in general, as well as on texts representing different time periods and genres (see further Section 3.1 for a description of the corpus used for evaluation). In this context, it should be noted that the sample from *Per Larssons dombok* that is part of the evaluation corpus is disjoint from the sample used for developing the normalisation rules.

3.2.1 Spelling in Different Time Periods and Genres

To get an idea of the differences in spelling between different time periods, the evaluation part of the Gender and Work corpus is divided into three (non-overlapping) subcorpora, representing time periods. All texts produced in the 16th century are grouped into the first subcorpus, whereas texts produced in the 17th century are grouped into the second subcorpus, and texts produced in the 18th century are grouped into the third subcorpus. Similarly for genres, the evaluation corpus is instead divided into two subcorpora, where all court records are grouped into the first subcorpus, and all church documents are grouped into the second subcorpus. These genre-specific subcorpora are non-

overlapping in relation to each other, but overlap with the first three time-specific subcorpora.

Table 3.2 shows the 10 most frequent spelling differences (at a character level) between the original historical text and the manually normalised version, in the five subcorpora. 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 the deletion of *h* (-*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. +*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 letter denoting the same sound. This can be illustrated by the change *w/v*, transforming for example *beswår* into *besvår* ('trouble').

16th century	17th century	18th century	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 3.2. Top 10 spelling differences at a character level, when comparing the original historical spelling to the manually normalised spelling in the Gender and Work corpus. 16th century texts are from the time period 1527–1593, 17th century texts from 1602–1689, and 18th century texts from 1728–1812. Court = Court records. Church = Church documents. Dbl = Double letter.

As seen from 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, but only the second most frequent transformation in 18th century text. This information indicates that the present normalisation rules may have a similar impact on all texts in the corpus, regardless of time period and genre.

3.2.2 Evaluation

Evaluation of the rule-based normalisation approach is performed in terms of *normalisation accuracy* and *error reduction*. Normalisation accuracy is defined as the percentage of tokens in the normalised text with a spelling identical to the manually modernised gold standard spelling. Error reduction is calculated by the following formula:

$$\text{error reduction} = \frac{\text{CorrectAfterNormalisation} - \text{CorrectBeforeNormalisation}}{\text{IncorrectBeforeNormalisation}}$$

CorrectAfterNormalisation = the percentage of tokens with an identical spelling to the modern version *after* the normalisation rules have been applied

CorrectBeforeNormalisation = the percentage of tokens with an identical spelling to the modern version *before* the normalisation rules have been applied

IncorrectBeforeNormalisation = the percentage of tokens differing in spelling from the modern version before the normalisation rules have been applied

The results are presented in Table 3.3, where the sample from *Per Larssons dombok* has been removed from the 17th century subcorpus as well as from the court records subcorpus, and evaluated separately. 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, a larger error reduction is achieved for this text than for the other subcorpora (38.6% as compared to the average 21.2%). This indicates that the normalisation rules are somewhat biased towards the text on which the rules were based. Nevertheless, normalisation has a positive effect on all subcorpora, regardless of time period and genre.

Text Type	Original	Normalised	Error Reduction
<i>Per Larssons dombok</i>	56.7%	73.4%	38.6%
16th Century	52.0%	64.2%	25.3%
17th Century	64.6%	72.5%	22.3%
18th Century	73.5%	77.6%	15.4%
Court Records	69.9%	75.6%	19.1%
Church Documents	54.9%	65.8%	24.2%
Average	63.0%	71.1%	21.2%

Table 3.3. Normalisation accuracy and error reduction for the rule-based approach when applied to different time periods and genres. *Original* = Proportion of words in the original text that are identical to the modern spelling in the gold standard. *Normalised* = Proportion of words in the normalised text that are identical to the modern spelling. *Per Larssons dombok* = sample from the same text as the (disjoint) sample used as a basis for rule development. *Average* is calculated over all time periods and genres, only excluding *Per Larssons dombok*.

The results show that the largest error reduction among the subcorpora, 25.3% as compared to the average 21.2%, is achieved for 16th century text. The fact that the oldest texts have the highest error reduction may not be very surprising, since the proportion of tokens with a modern spelling in the original, unnormalised text is generally lower in older texts, leaving more room for improvements. This means that even though the largest error reduction is achieved for the older texts, the younger texts still end up with a higher proportion of normalised tokens that are identical to the gold standard spelling.

Per Larssons dombok			
Text	Original	Normalised	Error Reduction
1638	56.7%	73.4%	38.6%
Court Records			
Text	Original	Normalised	Error Reduction
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	Original	Normalised	Error Reduction
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	65.8%	73.0%	21.8%

Table 3.4. Normalisation accuracy and error reduction for the rule-based approach when applied to different texts in the Gender and Work corpus. Original = Proportion of words in the original text that are identical to the modern spelling in the gold standard. Normalised = Proportion of words in the normalised text that are identical to the modern spelling. Per Larssons dombok = sample from the same text as the (disjoint) sample used as a basis for rule development. Average is calculated over all texts, only excluding Per Larssons dombok.

It could be observed that the results for church documents are similar to the results for 16th century text. This could be explained by the fact that the 16th

century part of the corpus contains only church documents. Table 3.4 provides more fine-grained results, distributed over all the single texts in the evaluation corpus. 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, as could be expected.

3.3 Levenshtein-based Normalisation

The Levenshtein distance gives an indication of the similarity between two strings, by computing the minimum number of characters that need to be inserted, deleted or substituted in order to transform one string into the other string [Levenshtein, 1966]. This could be summarised in the following formula, where $dist$ is the Levenshtein distance between two strings, with the first argument being the first string up to the i :th character, and the second argument being the second string up to the j :th character:

$$\begin{aligned}
 dist(0, 0) &= 0 \\
 dist(i, 0) &= i \\
 dist(0, j) &= j \\
 dist(i, j) &= \min \begin{cases} dist(i-1, j) + 1 & \text{deletion} \\ dist(i, j-1) + 1 & \text{insertion} \\ dist(i-1, j-1) + \begin{cases} 0 & \text{if } i=j \\ 1 & \text{otherwise} \end{cases} & \begin{matrix} \text{equality} \\ \text{substitution} \end{matrix} \end{cases}
 \end{aligned}$$

The Levenshtein distance is commonly used in spelling correction systems, where correction candidates are generated by edit distance comparisons between unknown words written by the user and word forms occurring in a dictionary. My Levenshtein-based approach to spelling normalisation builds on the same idea. Here, the normalisation process is viewed as a spelling correction problem, where the historical word form is treated as a misspelling that should be corrected to the most probable modern spelling. As seen from Figure 3.2, two steps are included in the normalisation process: *generation of normalisation candidates* and *candidate selection*.

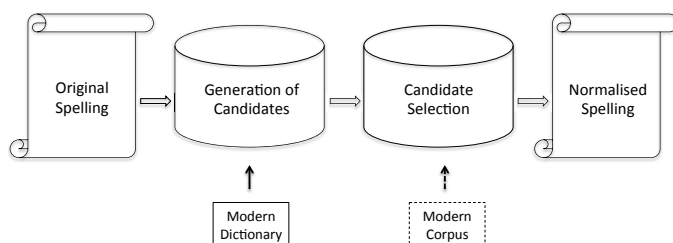


Figure 3.2. Overview of the Levenshtein-based spelling normalisation workflow, with optional resource (modern corpus) dotted.

In the generation step, each word form in the historical source text is compared to word forms present in a modern dictionary (or corpus). The dictionary entry with the lowest edit distance to the original word form is chosen for normalisation, provided that the distance is below a preset threshold value. If several dictionary entries share the same edit distance, all entries are extracted as possible normalisation candidates. In the selection step, one single candidate is to be selected as the final choice. In the default setting, a final candidate is randomly chosen from the list of highest-ranked candidates. If there is a modern language corpus available, the candidate with the highest frequency in the corpus is chosen. Only if none of the highest-ranked normalisation candidates are present in the corpus, or if there are several candidates with the same frequency distribution, a final candidate is randomly chosen.

The Levenshtein-based approach is similar to other approaches to spelling normalisation of historical text, as presented in Section 2.3.3. One difference is that while approaches such as the one presented by Bollmann et al. [2011] base their edit distance calculations on aligned parallel corpora of old spellings mapped to modern spellings, I compare the historical word form to modern dictionary entries. This means that no historical data is needed for training, which is an important aspect since this kind of data is lacking for many languages. However, if such a parallel corpus is available for the language at hand, this corpus may be used to further refine my Levenshtein approach by introducing weighted edit distance comparisons instead of the traditional Levenshtein comparisons (where all edits have a cost of 1).

In the following, I explore some settings for the Levenshtein-based approach, and how these settings affect normalisation accuracy and error reduction (calculated as described in Section 3.2.2). For this investigation, I use the training and evaluation corpora extracted from the Gender and Work corpus (further described in Section 3.1). As a modern dictionary for edit distance comparisons, I use SALDO (version 2.0), a lexical resource developed for present-day written Swedish, containing approximately 1,1 million word forms [Borin et al., 2008]. For corpus-based frequency statistics, I use the Stockholm Umeå Corpus (SUC, version 2.0) of text representative of the

Swedish language in the 1990s [Ejerhed and Källgren, 1997], containing approximately 1,2 million tokens.

3.3.1 String Length Restrictions

The basic idea of the Levenshtein-based normalisation approach is that the modern word form that is most similar to the historical word form is chosen for normalisation, based on edit distance comparisons towards a contemporary dictionary. It is however not a trivial decision what string similarity to regard as close enough for considering a normalisation candidate as valid. In a traditional spelling correction context, it has been shown that for the majority of the human-generated spelling errors, the misspelled word form is within one letter in length of the intended word form [Kukich, 1992]. If no empirical data on the correlation between historical and modern spelling is available, one could thus assume that a valid normalisation candidate should be maximally one character shorter or longer than the original word form. However, this might not be a suitable assumption, since spelling differences in historical text are due to other reasons than typos. To find out whether the traditional spelling correction assumptions hold also in the context of normalisation, I explore the characteristics of the manually normalised training corpus as regards differences in string length between the original word forms and their normalised counterparts.

Diff	Frequency	Original	Normalised	English
-7	1	<i>besluthninger</i>	<i>beslut</i>	‘decisions’
-6	3	<i>fadherbrodher</i>	<i>farbror</i>	‘uncle’
-5	7	<i>noghsambligha</i>	<i>nogsamma</i>	‘careful’
-4	127	<i>närvarellse</i>	<i>närvaro</i>	‘presence’
-3	308	<i>slechttenn</i>	<i>släkten</i>	‘the relatives’
-2	918	<i>kyrckian</i>	<i>kyrkan</i>	‘the church’
-1	4804	<i>aff</i>	<i>av</i>	‘of/off’
0	3326	<i>wara</i>	<i>vara</i>	‘be’
+1	1201	<i>til</i>	<i>till</i>	‘to’
+2	35	<i>sokn</i>	<i>socken</i>	‘parish’
+3	2	<i>tilbragte</i>	<i>tillbringade</i>	‘spent’
+5	3	<i>församhs</i>	<i>församlingens</i>	‘the congregation’s’

Table 3.5. Observed differences in string length between the historical word form and the manually normalised version in the training part of the Gender and Work corpus. Grey rows illustrate string length differences with a special focus in the experimental setup.

As shown in Table 3.5, the assumption that most errors do not influence string length by more than one character may still be regarded as appropriate even

in the context of normalisation. However, the proportion of tokens meeting this requirement is limited to approximately 86.9%. Instead, if we consider all cases where the normalised word form is at most one character longer down to four characters shorter than the original word form, approximately 99.5% of the tokens observed in the training corpus are covered. I therefore focus my string length experiments on varying the valid string length differences within this interval. It is interesting to note that the normalised word form often is shorter than the original historical word form, which could to a large extent be explained by diachronic simplifications in morphology and orthography.

Table 3.6 shows normalisation accuracy and error reduction when varying the threshold for valid string length differences between the original word form and its normalisation candidate(s). The results are also compared to the baseline, that is the unnormalised version of the evaluation corpus, where approximately 64.6% of the original tokens have a spelling identical to the manually normalised gold standard spelling. The proportion of tokens with a modern spelling increases to 77.0% after normalisation in the best Levenshtein setting, allowing for the normalised word to be one character longer or down to three characters shorter than the original historical word form. Hence, the general spelling correction assumption that the original string should not differ in length with more than one character from the intended word form is not optimal in this context. Increasing the threshold to allow for the normalised word form to be more than three characters shorter does however not have a noticeable effect on the results.

Approach	Accuracy	Error Reduction
Baseline	64.6%	n/a
Stringdiff +1 to -1	76.2%	32.7%
Stringdiff +1 to -2	76.9%	34.9%
Stringdiff +1 to -3	77.0%	35.0%
Stringdiff +1 to -4	77.0%	35.0%

Table 3.6. *Normalisation accuracy and error reduction for different normalisation settings when applying the Levenshtein-based normalisation approach to the evaluation part of the Gender and Work corpus. Baseline = Unnormalised version of the evaluation corpus. Stringdiff = Valid difference in string length between the original word form and its normalisation candidate(s).*

3.3.2 Edit Distance Restrictions

Apart from string length differences, the maximum edit distance allowed between the historical word form and the normalisation candidate(s) could also be taken into consideration. Similar to the spelling correction observation that string length normally does not differ with more than one character between a

misspelled word and the intended form, it has been shown that most misspelled word forms contain one single instance of insertion, deletion, or substitution, resulting in a maximum edit distance of 1 [Kukich, 1992]. As illustrated in Table 3.7, this assumption does not seem to be optimal in the context of spelling normalisation of historical text. If only normalisation candidates with a maximum Levenshtein distance of 1 as compared to the original word form are considered, approximately 55.8% of the token pairs in the training part of the Gender and Work corpus are covered. If instead an edit distance of up to and including 4 is considered, 98.8% of the observed entities in the corpus are covered. I therefore focus my experiments on varying the accepted edit distance between 1 and 4.

Distance	Frequency	Original	Normalised	English
1	5986	<i>sigh</i>	<i>sig</i>	‘oneself’
2	3008	<i>dher</i>	<i>där</i>	‘there’
3	1161	<i>blefwe</i>	<i>blev</i>	‘became’
4	455	<i>afwachta</i>	<i>avvakta</i>	‘await’
5	86	<i>öfwertalter</i>	<i>övertalad</i>	‘persuaded’
6	28	<i>söllfuermynthe</i>	<i>silvermynt</i>	‘silver coin’
7	10	<i>sielffuer</i>	<i>själv</i>	‘himself/herself’
8	1	<i>öffuergiffua</i>	<i>överge</i>	‘abandon’

Table 3.7. Observed edit distances between the historical word form and the manually normalised version in the training part of the Gender and Work corpus. Grey rows illustrate edit distances with a special focus in the experimental setup.

Table 3.8 presents my findings on varying the accepted edit distance between 1 and 4. For all settings in this experiment, the string length restrictions are set to the best-performing setting as presented in the previous section, that is letting the normalisation candidates differ in string length from the original word form by at most +1 to −3 characters. Again, the results show that it is not optimal to rely on findings from general spelling correction observations, stating that edit distance should not exceed a value of 1. For the evaluation part of the Gender and Work corpus, better results are achieved when increasing this threshold to 2, 3 or even 4. The differences between an edit distance of 2, 3, or 4 are however very small, and not statistically significant.

Approach	Accuracy	Error Reduction
Baseline	64.6%	n/a
Max distance 1	74.7%	28.5%
Max distance 2	76.7%	34.1%
Max distance 3	77.0%	34.9%
Max distance 4	77.0%	35.0%

Table 3.8. Normalisation accuracy and error reduction for different normalisation settings when applying the Levenshtein-based normalisation approach to the evaluation part of the Gender and Work corpus. Baseline = Unnormalised version of the evaluation corpus. Max distance = Maximum valid edit distance between the original word form and its normalisation candidate(s).

3.3.3 Weighted Edit Distance

When computing a traditional Levenshtein distance, all edit operations have a cost of 1. For example, the edit distance between the word forms *ryghtful* and *rightful* is 1, since one substitution of the letter *y* into *i* is required to turn the word form *ryghtful* into *rightful*. However, in the context of normalisation of historical text, some edits are more likely than others. For example, as historical texts to some degree are written in a spoken-language fashion, substituting the letter *y* for the letter *i* is more likely than substituting *y* for the phonologically more distant *r*. In the following, I explore the impact of assigning weights lower than 1 to edits occurring in the training corpus. Since this option is only possible when there is a training corpus of manually normalised word pairs available, it is interesting to see what effect this refinement has on the normalisation process.

For the calculation of weights, I first split the training corpus into a training part and a tuning part, by transferring 10% of the sentences to the tuning corpus, leaving 90% of the sentences in the training corpus. The training part of the corpus is used for extracting edits to consider, by automatically comparing the historical word forms to their modern spelling, using traditional Levenshtein edit distance comparisons. The edits extracted from the training corpus are then weighted based on their relative frequency in the tuning corpus, in accordance with the following formula:

$$\frac{\text{Frequency of Character Left Unchanged}}{\text{Frequency of Character}}$$

To illustrate, the deletion of the letter *h* is a commonly occurring edit operation when modernising historical Swedish word forms. This change is observed 202 times in the tuning part of the Gender and Work corpus. In the rest of the 318 cases where the letter *h* occurs, it is preserved in the modernised spelling. Thus, the weight for deletion of the letter *h* is calculated as:

$$\frac{318}{202+318} \approx 0.6115$$

Furthermore, edits may involve more than one character, as when comparing the Early Modern English spelling *personnes* to its modernised version *persons*. This would intuitively be regarded as deleting the French-alike ending *-ne*, rather than first deleting *-n* and then deleting *-e*. To handle this, I also include multi-character weights. The multi-character weights are calculated by the same formula as the single-character weights, and include the following operations:

- double deletion: *personnes* → *persons*
- double insertion: *strait* → *straight*
- single-to-double substitution: *jug*e → *jud*ge
- double-to-single substitution: *moost* → *most*

For all historical word forms in the training corpus that are not identical in the modern spelling, all possible single-character edits as well as multi-character edits are counted for weighting. Hence, the historical word form *personnes*, mapped to the modern spelling *persons*, will yield weights for double deletion of *-ne*, as illustrated above, but also for single deletion of *-n* and single deletion of *-e*, as well as double-to-single substitution of *-nn* to *-n* or *-ne* to *-n*.

Normalisation results for weighted edit distance are presented in Table 3.9. For these experiments, the string length and edit distance restrictions are set to the best-performing settings as presented in the previous sections, that is letting the normalisation candidates differ in string length from the original word form by at most +1 to −3 characters, with a maximum edit distance of 4.

Approach	Accuracy	Error Reduction
Baseline	64.6%	n/a
No weights	77.0%	35.0%
Single-character weights	78.7%	39.9%
Multi-character weights	79.1%	40.9%

Table 3.9. Normalisation accuracy and error reduction for different normalisation settings when applying the Levenshtein-based normalisation approach to the evaluation part of the Gender and Work corpus. Baseline = Unnormalised version of the evaluation corpus.

As seen from the results, the inclusion of (single-character) weights leads to an increase in normalisation accuracy from 77.0% to 78.7%, as compared to normalisation using standard Levenshtein comparisons. When supplementing single-character weights with multi-character weights, normalisation accuracy

increases further to 79.1%. It is likely that the inclusion of weights would have an even larger impact on normalisation accuracy if more training data was available.

3.3.4 Compound Splitting

The Levenshtein approach to spelling normalisation is based on comparison of historical word forms to similar word forms found in a modern dictionary. Since the Swedish language has a high degree of compounds, some of the intended word forms will inevitably not be found in the dictionary, even if the word could perfectly well be used in contemporary Swedish. Szymne and Holmqvist [2008] showed that in a corpus of contemporary Swedish text, approximately 37% of all the words with a length of 12 characters or longer were compounds (and approximately 5% of the shorter words). This was calculated on modern Swedish European Parliament text, but might still be indicative of the frequencies of compounds in historical text as well.

To deal with the compounding issue, I include a compound splitter developed by Szymne [2008]. Based on statistics of word frequencies in a training corpus, the compound splitter splits presumptive compounds into smaller parts matching word forms occurring in the training corpus. In the context of spelling normalisation, the historical spelling in the training part of the Gender and Work corpus is used for training the compound splitter. During normalisation, all word forms for which no appropriate normalisation candidate is found by the ordinary edit distance calculations, are run through the compound splitter. If the splitter is able to split the word into compound parts, each part is processed separately by the normalisation program, and the resulting normalisation candidates are merged into a final normalisation candidate. For example, a word like *krigztienst* will be split into *krigz* (normalised as *krigs* ‘war’) and *tienst* (normalised as *tjänst* ‘duty’). The two normalised versions *krigs* and *tjänst* are then merged into the final normalisation candidate *krigstjänst* (‘military service’).

Adding the compound splitter to the normalisation process has a small but positive effect on the normalisation process. The normalisation accuracy is still 79.1%, as for the best Levenshtein setting presented in the previous section. At a token level however, 26,997 tokens in the normalised text are identical to the manually modernised spelling in the original setting, as compared to 27,013 identically spelled tokens in the compound setting. One reason for the barely noticeable improvement may be the rather small training corpus. To illustrate, Szymne [2008] experimented on splitting German compounds based on a training corpus of 1,467,291 sentences, as compared to the historical training corpus of only 600 sentences.

3.4 Memory-based Normalisation

The core component of the memory-based approach to normalisation is a parallel training corpus of token pairs with historical word forms mapped to their manually modernised spelling. This training corpus is used as a memory in the normalisation process, providing information on how this word form has been normalised in previously seen texts. Whenever a token is encountered that also occurs in the training data, the most frequent modern spelling associated with that token in the training corpus is chosen for normalisation. Tokens that are not found in the training corpus are left unchanged. The whole normalisation workflow is illustrated in Figure 3.3.

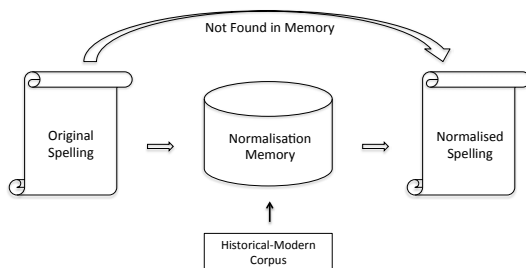


Figure 3.3. Overview of the memory-based spelling normalisation workflow.

The memory-based approach is trained and evaluated on the same Gender and Work corpus as in the evaluation of the rule-based normalisation approach and the Levenshtein-based normalisation approach (see further Section 3.1 for a description of this corpus). As stated in Section 1.1, one of my goals is to develop normalisation methods that are applicable even in cases where only small amounts of training data is available. Therefore I also split the training corpus into smaller parts, to enable the evaluation of normalisation accuracy with different sizes of the training corpus. The results are given in Figure 3.4.

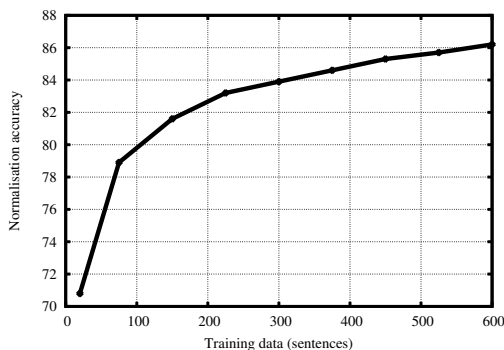


Figure 3.4. Normalisation accuracy for the memory-based approach, varying the size of the training corpus.

The smallest subpart of the corpus contains exactly the same sentences as were used for developing the hand-crafted rules for the rule-based method, that is the first 20 sentences of the court records text *Per Larssons dombok*. For the memory-based approach, this resulted in a normalisation accuracy of 70.8%, as compared to 73.0% for the rule-based approach. It was expected that the rule-based approach would perform better in this setting, since the data set is rather small, and the memory-based approach is dependent on the whole word to occur in the training data, whereas the rule-based approach operates on changes at a character level. At the next data point, the training corpus contains instead five sentences from each subcorpus (i.e. in total 75 sentences), improving normalisation accuracy to 78.9%. For each of the succeeding data points in the figure, five new sentences from each subcorpus were added to the training data. As could be expected, the more data that is included in the training corpus, the higher the accuracy score. For the full set of 600 sentences, a normalisation accuracy of 86.2% is achieved, as compared to 73.0% for the rule-based approach, and 79.1% in the best-performing Levenshtein setting. Provided that a historical-modern parallel corpus of reasonable size is available for a particular language, it is thus likely that the rather simple data-driven memory-based approach would outperform the rule-based approaches in terms of normalisation accuracy.

3.5 SMT-based Normalisation

In the SMT-based approach, spelling normalisation is treated as a translation task, which could be solved using statistical machine translation (SMT) techniques. The fundamental task of an SMT system is to find the best target language translation of a given source language sentence, in accordance with a probabilistic model. SMT models are usually trained on large amounts of example data, in the form of word-aligned parallel corpora for training a *translation model*, as well as monolingual target language corpora for training a *language model*.

The translation model captures the likelihood that certain units in the target language are translations of a string in the source language. Common features that are used to express this likelihood are based on co-occurrence frequencies in word-aligned parallel corpora. Parallel corpora consist of texts that are available in two (or more) languages. For the corpus to be useful for machine translation purposes, the parallel texts are aligned, mapping sentences, words and word sequences in the source language to the corresponding units in the target language(s). In the sentence alignment step, each sentence in the source language is aligned to the corresponding target language sentence(s). The alignments may be:

tion by modeling the probabilities that the candidate translation strings would occur in the target language [Koehn, 2010].

In contrast to traditional machine translation, the normalisation task should address changes in spelling rather than full translation of words and phrases. Therefore, translation in the spelling normalisation context is performed at a character level. The basic idea of character-level SMT is that phrases are modeled as character sequences instead of word sequences. Translation models are then trained on character-aligned parallel corpora, and language models on character N-grams. With this kind of training data, word alignment can be modeled in the same way as sentence alignment is modeled otherwise, whereas character alignment is performed using standard word alignment techniques [Tiedemann, 2009].

The parallel corpus used for the translation model in the spelling normalisation setup consists of texts available in their original, historical spelling as the source language, and their manually modernised spelling as the target language. Since the target language differs only in spelling as compared to the source text, without word reordering or deletion and insertion of words, a one-to-one correspondence between the words in the source text and the words in the target text can be assumed. Thus, word alignment is a rather trivial task, which can be performed using standard sentence alignment techniques. The next step is character alignment, which is performed using standard word alignment techniques, with input data formatted into one token on each line, with a blank space separating each character, as in the example below:

T o	T o
t h e	t h e
m o o s t	m o s t
n o b l e	n o b l e
&	a n d
W o r t h i e s t	w o r t h i e s t
L o r d e s	L o r d s
m o o s t	m o s t
r y g h t f u l	r i g h t f u l
&	a n d
w y s e s t	w i s e s t
c o n s e i l l e	c o u n c i l

The characters are aligned using the same techniques as would in the more traditional SMT setup be used for word alignment. Figure 3.6 illustrates the character alignment step in the context of spelling normalisation, aligning the Middle English spelling *moost* to its modern spelling *most*.

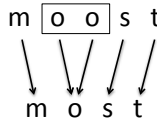


Figure 3.6. Character alignment in the SMT-based approach to spelling normalisation.

It has previously been shown that character-based SMT is useful for transliteration and translation between closely related languages [Matthews, 2007, Vilar et al., 2007, Tiedemann and Nabende, 2009, Tiedemann, 2009, Nakov and Tiedemann, 2012]. In a way, historical spelling and modern spelling could be seen as closely related languages, which is why I follow these ideas in the SMT-based approach to spelling normalisation. Furthermore, Nakov and Tiedemann [2012] have shown that small parallel training corpora are sufficient for reasonable performance of character-based SMT systems, which is desirable in this context, due to the general shortage of parallel corpora with historical spelling aligned to modern spelling. Language models on the other hand can be trained on larger monolingual corpora, which are often available for the modern language, and higher orders in terms of N-gram size can be used to ensure fluent and grammatically correct output.

In my SMT-based normalisation setup, I perform character-based SMT with a phrase-based translation model, using the SMT engine Moses with all its standard components [Koehn et al., 2007], and IRSTLM for language modeling [Federico et al., 2008]. An overview of the SMT setup for spelling normalisation of historical text is given in Figure 3.7.

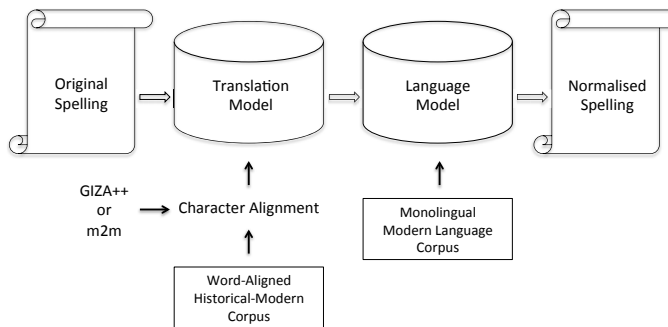


Figure 3.7. Overview of the SMT-based spelling normalisation workflow.

As previously mentioned, it is crucial for the final translation quality that the parallel corpus used as a basis for the translation model is properly aligned. In my experiments, I try two different character alignment techniques:

1. The word alignment toolkit GIZA++ [Och and Ney, 2003], implementing the IBM models commonly used in SMT [Brown et al., 1993].
2. A weighted finite state transducer implemented in the m2m-aligner [Jiampoamarn et al., 2007], where the transducer models are based on context-independent single-character and multi-character edit operations.

An example of character alignment using the m2m-aligner is given in Figure 3.8, illustrating the alignment of the Middle English spellings *peple*, *owre*, *ded* and *moost* to the modern versions *people*, *our*, *dead* and *most* respectively. In this example, the ϵ symbol denotes empty alignments, that is insertions and deletions. Hence, the ϵ symbol in the source word *peple* denotes the insertion of *o* in the target word *people*. Likewise, the ϵ symbol in the target word *our* denotes the deletion of *e* as compared to the source word *owre*. Two-to-one and one-to-two alignments are also possible, as in the case of the alignment of *moost* to *most*, where the colon denotes that the succeeding two *o*:s in the source word correspond to one single *o* in the target word. Similarly for the alignment between *ded* and *dead*, the colon in the target word *dead* means that the *e* in the source word corresponds to both *e* and *a* in the target word.

```

p|e| $\epsilon$ |p|l|e|   o|w|r|e|   m|o:o|s|t|   d| e |d|
p|e|o|p|l|e|   o|u|r| $\epsilon$ |   m| o |s|t|   d|e:a|d|

```

Figure 3.8. Character level alignment using the m2m-aligner.

In the following experiments, I explore how different characteristics of the training data affects the final normalisation quality in the SMT-based setting. More specifically, the following experiments are conducted and evaluated:

1. Different techniques for word alignment

As described above, a word-aligned historical-modern corpus is used as input for estimating the translation model in the SMT-based normalisation setup. Manually word-aligned corpora of this kind are available for some languages, where historical texts have typically been stored in an XML format, with the normalised token assigned as a feature to the original, historical word form. Sometimes however, parallel historical-modern corpora are available in an unaligned format only, with no explicit mapping between the original tokens and their normalised forms. Even though spelling normalisation should only concern differences at a character level, and generally does not affect the word order, word alignment in this context is less trivial than could be expected. In some cases, several word forms in the historical spelling are mapped to a single word form in the normalised version, or vice versa. This could be exemplified by the normalisation of the Icelandic proper name *Snorri Stvrlo son* (literally meaning ‘Snorri, the son of Stvrlo’) into the modern form of the name *Snorri Sturluson*. There are also cases where the

person who normalised the text has ‘corrected’ the normalised version by inserting missing words or deleting superfluous words etc. Due to these discrepancies between the source and target texts, it is interesting to explore the effect of using manually word-aligned training data in the normalisation process, as compared to using automatically word-aligned corpora.

2. Different sizes of the parallel corpus

In traditional SMT, large amounts of parallel training data are generally needed to achieve a high translation quality, due to the large set of possible words and word sequences occurring in a language. In the context of character-based SMT on the other hand, the set of possible characters and character sequences in a language is substantially smaller. Therefore, it could be expected that less training data would be needed for this task. This is an important aspect for the normalisation task, since there is limited access to parallel corpora of texts available in both their original, historical spelling and their manually modernised version. I therefore explore the impact of varying the size of the parallel corpus used for estimating the training model.

3. Different sizes and genres of the target language corpus

In the SMT-based spelling normalisation setup, a corpus of contemporary texts is used for estimating the language model. Even though such corpora are available for many languages, the size and contents of these corpora differ between the languages. For some languages there are large, balanced corpora available with texts from a variety of genres. For other languages, there may be only a smaller corpus available, possibly containing one or two genres only. I therefore experiment on varying the size and genres included in the contemporary corpus used for language modeling, with the aim of exploring whether the SMT-based normalisation approach is applicable also to languages for which there is limited access to contemporary corpora.

4. Different tools and settings for character alignment

As previously discussed, the correctness of the alignments in the parallel corpus used as a basis for translation modeling is crucial for the final normalisation quality. I therefore try different tools and settings for the character alignment step, in order to find the best-performing settings for the specific task of spelling normalisation.

In order to perform the experiments on different techniques for word alignment, I need access to a gold standard of manually normalised and word-aligned historical text. I also need access to the historical and modern versions of the same text in its original, unaligned format, to be used as input data to the automatic word alignment tools. The Swedish Gender and Work data is not suitable for this purpose, since the manual normalisation of this

corpus has been performed on tokenised data only, resulting in a manually validated word-aligned normalisation which could not be used as input to the automatic alignment tools. For Icelandic on the other hand, I have access to a historical-modern parallel text from the 14th century Icelandic saga *Snorri Sturluson’s Edda* (the Uppsala version DG11, [Palsson, 2012]), available both in a document-level normalisation format and in the form of manually validated word alignments. From this parallel corpus of 33,888 token pairs, every 10th sentence was extracted to a tuning corpus, and the rest of the sentences were stored as a training corpus, resulting in a training set of 30,451 token pairs and a tuning set of 3,437 token pairs. For language modeling, I use all tokens occurring 100 times or more in the *Tagged Icelandic Corpus of Contemporary Icelandic texts*, MÍM [Helgadóttir et al., 2012]. The rather high frequency threshold is chosen due to a considerable amount of noisy corpus data, especially in the texts extracted from blogs and other websites. In total, this subset of the corpus contains 21,613,551 tokens, with texts distributed over 12 genres, including for example newspaper text, blog text, parliamentary speeches and university essays. Evaluation is performed on a subset of *Ectors saga* from the 15th century [Loth, 1962]. This text contains 20,811 tokens and is part of the Icelandic Parsed Historical Corpus, IcePaHC [Rögnvaldsson et al., 2012]. The Icelandic corpus distribution is given in Table 3.10.

	Name	Time Period	Tokens
Training	Snorri Sturluson’s Edda	14th century	33,888
Evaluation	Ectors saga	15th century	20,811

Table 3.10. Number of tokens in the training and evaluation parts of the Icelandic corpus.

The experiments on varying the size of the parallel corpus, as well as the size and genre of the contemporary corpus used for language modeling, are performed on the same Icelandic data sets, whereas the experiments on using different tools and settings for character alignment are performed both for Icelandic and for Swedish. For Swedish, I then use the same training, tuning and test sets from the Gender and Work corpus as were used in the previous experiments on Levenshtein-based and memory-based normalisation. The modern corpus used for language modeling in the Swedish setting is the Stockholm Umeå Corpus (SUC, version 2.0) of text representative of the Swedish language in the 1990s [Ejerhed and Källgren, 1997], containing approximately 1,2 million tokens.

3.5.1 Different Techniques for Word Alignment

As discussed above, a word-aligned training corpus of historical tokens mapped to their modern spelling is needed for performing spelling normalisation using

character-based SMT techniques. Even in cases where a historical text has been manually normalised to a modern spelling, the parallel texts are not always aligned at a word level. Since manual word alignment is time-consuming and requires language-specific knowledge, it is therefore interesting to investigate the impact of performing SMT normalisation based on an automatically word-aligned training corpus as compared to using manually word-aligned training data.

For the word alignment task, I use *hunalign*, which is in fact a sentence aligner rather than a word aligner [Varga et al., 2005]. For the case of aligning historical and modern spelling however, monotonic alignments without reordering are assumed, with a strong correlation between the length of the original spelling and the modern spelling. These features make it suitable to use common length-based and lexical matching-based sentence alignment algorithms, like the one implemented in *hunalign*. I try four different ways of running *hunalign* for this specific task:

1. **no split**

Input data is one token on each line, with empty lines denoting sentence boundaries.

2. **no split +realign**

Same as “no split”, but with the additional flag *-realign*, in which the aligner is run in three phases, heuristically building a dictionary based on the identified sentence pairs (in this case word pairs).

3. **split**

Same as “no split”, but with whitespace separating all characters.

4. **split +realign**

Same as “split”, but with the additional *-realign* flag, as described above. Characters are treated as tokens in the lexical matching.

Automatic alignment will inevitably introduce noise in the training data. Since the modern version is stated to be a modernisation of spelling, not including syntactic normalisation, the word alignment task is easier than in an ordinary translation setting, but still not trivial, as discussed above. To evaluate the impact of the different alignment methods on the normalisation results, I ran normalisation experiments based on the GIZA++ unigram setting for character alignment, with training data automatically generated from the four alignment methods described above. The results are presented in Table 3.11, with comparison to the baseline proportion of modern spellings in the original, unnormalised text, and to the upper-bound results achieved for manually word-aligned training data.

Alignment Model	Normalisation Accuracy
baseline	64.8%
no split	78.8%
no split +realign	78.8%
split	79.1%
split +realign	79.2%
manual alignment	83.9%

Table 3.11. *Normalisation accuracy for the SMT-based approach to spelling normalisation, using different methods for word alignment of the Icelandic historical-modern training data. Baseline = Percentage of tokens in the unnormalised text with a spelling identical to the modern spelling.*

Approximately two thirds (64.8%) of the original tokens in the evaluation corpus already have a spelling that is identical to the modern spelling. Normalisation based on the best automatic word alignment setting (SPLIT +REALIGN) significantly increases this proportion to 79.2%. It is also clear that splitting the words into their separate characters has a positive effect on the alignment results. With the SPLIT setting, normalisation accuracy is 79.1% as compared to 78.8% for the NO SPLIT setting. The REALIGN flag however, does not seem to have a noticeable effect on the normalisation results. It is also worth mentioning that normalisation accuracy for automatically generated training data is fairly close to the accuracy achieved for manually aligned training data; 79.2% in the best hunalign setting as compared to 83.9% for manually aligned data. Hence, if no word-aligned historical-modern data is available, automatic alignment techniques may successfully be used to automatically create such a parallel corpus.

3.5.2 Different Sizes of the Parallel Corpus

Since one of the goals of my approach to NLP for historical text is to develop solutions that are applicable even when only small amounts of historical data is available, I experiment on varying the size of the historical-modern parallel corpus used for estimating the translation model in the SMT-based normalisation approach. Figure 3.9 presents the results for Icelandic (using manually aligned training data).

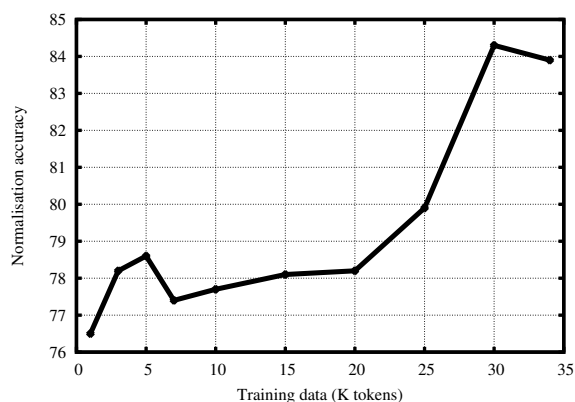


Figure 3.9. Normalisation accuracy for the SMT-based approach to spelling normalisation, varying the size of the Icelandic historical-modern corpus used for estimating the translation model.

In these experiments, all parameters except the size of the training data are kept unchanged. As expected, the more training data, the better normalisation results (in general). However, with only 1,000 token pairs, a normalisation accuracy of 76.5% is achieved, as compared to 83.9% for the entire corpus. This demonstrates that fairly good normalisation results can be achieved also in cases where there is limited access to parallel historical-modern text. The drop in accuracy for sizes above 5,000 tokens, but below 20,000 tokens, further indicates that not only the quantity but also the quality of the data is important to consider. Introducing noise in the training data is particularly unfortunate for small data sets.

3.5.3 Different Sizes and Genres of the Target Language Corpus

Apart from the size of the training data, another aim of my approach to NLP for historical text is that it should be useful for any language where there is a basic set of language resources and tools available for the modern version of the language. Such a set of language resources may contain corpora of varying sizes, sometimes including several genres and sometimes including for example newspaper text only. To test the applicability of the SMT-based normalisation approach to different sizes and genres included in the language model data, I perform experiments on varying the size of the target language corpus, from 1 million tokens as a minimum to including all $\sim 21,000,000$ tokens in the contemporary Icelandic corpus. Furthermore, I compare the results for including newspaper text only to the results for including all the genres of the corpus in the language model. In these experiments, all parameters except the size of the language model data are kept unchanged. The results are presented in Figure 3.10.

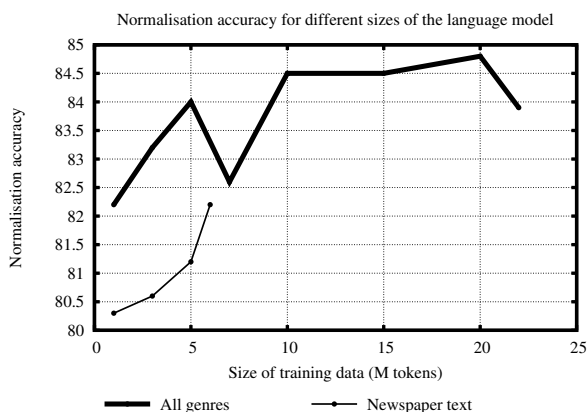


Figure 3.10. Normalisation accuracy for the SMT-based approach to spelling normalisation, varying the size and genres included in the contemporary Icelandic corpus used for language modeling.

As seen from the results, including a corpus of different genres in the language model shows better results than using newspaper text only. The difference is however not very large between the two types of corpora. For 5 million words of newspaper text, a normalisation accuracy of 81.2% is achieved, as compared to 84.0% for the same amount of tokens in the sampled corpus. It is also worth mentioning that whereas the newspaper corpus shows the expected increase in normalisation accuracy when more data is added, this relation is not as clear-cut for the sampled corpus where the addition of more data in some cases leads to a drop in normalisation accuracy. This could be due to larger variation in the sampled training data, meaning that adding new data sometimes distort the language model. In fact, including 10, 15 or 20 million tokens in the language model yields higher normalisation accuracy than including the full corpus of ~21,000,000 tokens. Also note that including only 1 million (sampled) tokens results in a normalisation accuracy of 82.2%, which is already close to the 83.9% achieved when including the full corpus in the language model.

3.5.4 Different Tools and Settings for Character Alignment

In the previous experiments on word alignment techniques and different sizes of the training data, the character alignment step in the translation modeling phase is performed using the basic GIZA++ unigram setting. In this section, I compare the normalisation results when using GIZA++ for character alignment, as compared to using the m2m aligner. I also try unigram-based and bigram-based alignment for both methods. The results for Icelandic and Swedish are summarised in Table 3.12.

Alignment Setting	Normalisation Accuracy	
	Swedish	Icelandic
baseline	64.6%	64.8%
GIZA++ unigram	92.9%	83.9%
GIZA++ bigram	92.5%	83.5%
m2m unigram	92.3%	81.0%
m2m bigram	92.2%	81.4%

Table 3.12. *Normalisation accuracy for the SMT-based approach to spelling normalisation, using different character alignment models. Baseline = Percentage of tokens in the unnormalised text with a spelling identical to the modern spelling.*

In both the Swedish and the Icelandic evaluation corpus, approximately two thirds of the original tokens already have a spelling that is identical to the modern spelling. For both languages, the GIZA++ unigram setting results in the highest normalisation accuracy. Using more informative bigrams instead of unigrams leads to a drop in normalisation accuracy for most settings. Likewise, the m2m aligner is slightly less successful for the normalisation task at hand, as compared to the GIZA++ unigram model. The differences between some of the models are however small, and not statistically significant.

4. Multilingual Evaluation

In Chapter 3, four approaches to spelling normalisation were presented and evaluated on the Swedish Gender and Work corpus (and to some extent on Icelandic data). In this chapter I present an evaluation of the applicability of three of these normalisation methods to five different languages: English, German, Hungarian, Icelandic, and Swedish. The aim is to explore whether the proposed methods are language-independent enough to be applicable to other languages as well, and if certain methods work better for specific languages or if the same normalisation approach always yields the highest normalisation accuracy, regardless of language.

Due to its language-specific nature, the rule-based normalisation approach is excluded from the multilingual evaluation. Instead, a combination of the Levenshtein-based approach and the memory-based approach is included, in addition to the methods presented in Chapter 3. In the combined method, the normalisation memory is consulted before possible normalisation candidates are generated based on Levenshtein distance. If the historical word form is present in the memory, the most frequent normalised form found in the memory is chosen. Only word forms that are not found in the memory are normalised by means of Levenshtein comparisons. In summary, the evaluation thus covers the following four normalisation methods:

1. Levenshtein-based normalisation
2. memory-based normalisation
3. combination of Levenshtein-based and memory-based normalisation
4. SMT-based normalisation

4.1 Experimental Setup

For the spelling normalisation experiments, five data sets are required:

1. A training corpus of historical data available both in its original spelling and with a manually modernised spelling.
2. A tuning corpus of historical data available both in its original spelling and with a manually modernised spelling.

3. An evaluation corpus of historical data available both in its original spelling and with a manually modernised spelling.
4. A contemporary dictionary used as a basis for edit distance comparisons in the Levenshtein-based spelling normalisation approach. If no dictionary is available, a contemporary corpus may be used instead.
5. A contemporary corpus used for frequency calculations in the Levenshtein-based spelling normalisation approach, and as a language model in the SMT-based normalisation approach.

For convenience, the notions of training, tuning and evaluation corpora are used, which are well-known concepts within SMT. These data sets are created by extracting every 9th sentence from the total corpus to the tuning corpus, and every 10th sentence to the evaluation corpus, whereas the rest of the sentences compose the training corpus. The only exception is Swedish, where I use the same training, tuning and test corpora as were created for the normalisation experiments described in Chapter 3. In the memory-based approach, there is in fact no distinction between training and tuning corpora, since both data sets are combined in the lookup process. As for the Levenshtein edit distance approach, the training corpus is used for extracting single-character and multi-character edits by comparing the historical word forms to their modern spelling. The edits extracted from the training corpus are then weighted based on their relative frequency in the tuning corpus. The tuning corpus is also used for setting a threshold for which maximum edit distance to allow between the original word form and its normalisation candidate(s). This threshold is calculated by the following formula (where 1.96 times the standard deviation is added to to give a 95% confidence interval:

$$\text{average edit distance} + (1.96 * \text{standard deviation})$$

Even though the Levenshtein normalisation results presented in Section 3.3.2 suggests a maximum edit distance of 4, those experiments were performed for Swedish only. Since the optimal threshold probably varies between different languages, this formula is used to generate language-specific maximum edit distance thresholds in the multilingual setting.

The historical texts used for training, tuning and evaluation need to be available both in their original, historical spelling and in a manually modernised and validated spelling. A modern translation of a historical text is generally not usable, since word order and sentence structure have to remain the same to enable training and evaluation of the proposed methods. The access to such data is very limited, meaning that the data sets used in the experiments vary in size, genres and time periods between the languages. The evaluation is however not geared towards a comparison between the languages, but rather

to test the generalisability of the methods to different languages, time periods and genres. In the following, the data sets for each language are described in more detail.

4.1.1 English

For training, tuning and evaluation in the English experiments, I use the *Innsbruck Corpus of English Letters* (version 2.1), a manually normalised collection of letters from the period 1386–1698. This corpus is part of the *Innsbruck Computer Archive of Machine-Readable English Texts*, ICAMET [Markus, 1999]. A subset of the British National Corpus (BNC) is used as the single modern language resource. Table 4.1 presents in more detail the data sets used in the English experiments.

Resource	Data	Tokens	Types
Historical-Modern Training Corpus	ICAMET	148,852	18,267
Historical-Modern Tuning Corpus	ICAMET	16,461	4,391
Historical-Modern Evaluation Corpus	ICAMET	17,791	4,573
Contemporary Dictionary Resource	BNC	2,088,680	69,153
Contemporary Corpus Resource	BNC	2,088,680	69,153

Table 4.1. *Language resources for English.*

4.1.2 German

For training, tuning and evaluation in the German experiments, I use a manually normalised subset of the *GerManC* corpus of German texts from the period 1650–1800 [Scheible et al., 2011a]. This subset contains 22 texts from the period 1659–1780, within the genres of drama, newspaper text, letters, sermons, narrative prose, humanities, science och legal documents. The German *Parole* corpus is used as the single modern language resource [Teubert, 2003]. Table 4.2 presents in more detail the data sets used in the German experiments.

Resource	Data	Tokens	Types
Historical-Modern Training Corpus	GerManC	39,887	9,055
Historical-Modern Tuning Corpus	GerManC	5,418	2,056
Historical-Modern Evaluation Corpus	GerManC	5,005	1,966
Contemporary Dictionary Resource	Parole	18,662,243	662,510
Contemporary Corpus Resource	Parole	18,662,243	662,510

Table 4.2. *Language resources for German.*

4.1.3 Hungarian

For training, tuning and evaluation in the Hungarian experiments, I use a collection of manually normalised codices from the *Hungarian Generative Diachronic Syntax* project, HGDS [Simon, 2014], in total 11 codices from the time period 1440–1541. The Szeged Treebank is used as the single modern language resource [Csendes et al., 2005]. Table 4.3 presents in more detail the data sets used in the Hungarian experiments.

Resource	Data	Tokens	Types
Historical-Modern Training Corpus	HGDS	137,669	45,529
Historical-Modern Tuning Corpus	HGDS	17 181	8 827
Historical-Modern Evaluation Corpus	HGDS	17,214	8,798
Contemporary Dictionary Resource	Szeged	1,257,089	144,248
Contemporary Corpus Resource	Szeged	1,257,089	144,248

Table 4.3. Language resources for Hungarian.

4.1.4 Icelandic

For training, tuning and evaluation in the Icelandic experiments, I use a manually normalised subset of the *Icelandic Parsed Historical Corpus* (IcePaHC, version 0.9), a manually tagged and parsed diachronic corpus of texts from the time period 1150–2008 [Rögnvaldsson et al., 2012]. This subset contains four texts from the 15th century: three sagas (*Vilhjálms saga*, *Jarlmann’s saga*, and *Ector’s saga*) and one narrative-religious text (*Miðaldaævintýri*). The contemporary dictionary resource is a combination of *Beygingarlýsing Íslensks Nútímamáls*, BÍN (a database of modern Icelandic inflectional forms [Bjarnadóttir, 2012]), and all tokens occurring 100 times or more in the *Tagged Icelandic Corpus of Contemporary Icelandic texts*, MÍM [Helgadóttir et al., 2012].¹ The high frequency threshold of 100 is chosen due to a considerable amount of noisy corpus data, especially in the texts extracted from blogs and other websites. The tokens occurring 100 times or more in the MÍM corpus are also used as the contemporary corpus resource. Table 4.4 presents in more detail the data sets used in the Icelandic experiments.

¹The BÍN database alone is not sufficient for Levenshtein calculations, since it only contains content words.

Resource	Data	Tokens	Types
Historical-Modern Training Corpus	IcePaHC	52,440	9,748
Historical-Modern Tuning Corpus	IcePaHC	6,443	2,270
Historical-Modern Evaluation Corpus	IcePaHC	6,384	2,244
Contemporary Dictionary Resource	BÍN+MÍM	27,224,798	2,820,623
Contemporary Corpus Resource	MÍM	21,339,384	9,461

Table 4.4. *Language resources for Icelandic.*

4.1.5 Swedish

For training, tuning and evaluation in the Swedish experiments, I use balanced subsets of the Gender and Work corpus (GaW) of court records and church documents from the time period 1527–1812 [Ågren et al., 2011] (see further Section 3.1 for a description of the Gender and Work corpus). The dictionary resource used is SALDO (version 2.0), a lexical resource developed for present-day written Swedish [Borin et al., 2008]. The contemporary corpus resource used is the Stockholm Umeå corpus (SUC, version 2.0) of text representative of the Swedish language in the 1990s [Ejerhed and Källgren, 1997]. Table 4.5 presents in more detail the data sets used in the Swedish experiments.

Resource	Data	Tokens	Types
Historical-Modern Training Corpus	GaW	28,237	7,925
Historical-Modern Tuning Corpus	GaW	2,590	1,260
Historical-Modern Evaluation Corpus	GaW	33,544	8,859
Contemporary Dictionary Resource	SALDO	1,110,731	723,138
Contemporary Corpus Resource	SUC	1,166,593	97,670

Table 4.5. *Language resources for Swedish.*

4.2 Results

Table 4.6 presents the results for different languages and normalisation methods, given in terms of *normalisation accuracy* (i.e. the percentage of tokens in the normalised text with a spelling identical to the manually modernised gold standard) and *character error rate (CER)*, providing a more precise estimation of the similarity between the normalised token and the gold standard version at a character level. For the Levenshtein-based approach, single-character and multi-character weights are included, as described in Section 3.3. No compound splitter is included, due to its language-specific nature. In the SMT-based approach, both GIZA++ and the m2m aligner are run with standard word alignment models for character unigrams (un) and bigrams (bi).

	English		German		Hungarian		Icelandic		Swedish	
	Acc	CER	Acc	CER	Acc	CER	Acc	CER	Acc	CER
baseline	75.8	0.26	84.4	0.16	17.1	0.85	50.5	0.51	64.6	0.36
Lev	82.9	0.19	87.3	0.13	31.7	0.71	67.3	0.35	79.4	0.22
memory	91.7	0.20	94.6	0.26	75.0	0.30	81.7	0.25	86.2	0.27
Lev+memory	92.9	0.09	95.1	0.06	76.4	0.35	84.6	0.19	90.8	0.10
GIZA++ un	94.3	0.07	96.6	0.04	79.9	0.21	71.8	0.30	92.9	0.07
GIZA++ bi	92.4	0.09	95.5	0.05	80.1	0.21	71.5	0.30	92.5	0.08
m2m un	90.6	0.11	96.0	0.04	79.4	0.21	71.2	0.31	92.3	0.08
m2m bi	88.0	0.14	95.6	0.05	79.5	0.21	71.5	0.30	92.2	0.08

Table 4.6. Normalisation results given in normalisation accuracy (Acc) and character error rate (CER).

Table 4.7 summarises the results in terms of *Precision (Pre)*, *Recall (Rec)* and *F-score (F)* for the Levenshtein-based approach, the memory-based approach, the combined Levenshtein and memory-based approach, and the best-performing SMT-based approach.

	memory			Levenshtein			Lev+memory			SMT		
	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F
English	93.6	97.8	95.7	92.7	88.6	90.7	97.4	95.2	96.3	98.2	95.9	97.0
German	95.0	99.6	97.2	91.0	95.6	93.2	97.3	97.7	97.5	98.7	97.9	98.3
Hungarian	77.4	96.0	85.7	68.0	37.3	48.2	96.2	78.8	86.7	98.3	81.3	89.0
Icelandic	89.3	90.6	89.9	85.4	76.1	80.5	95.6	88.0	91.7	82.0	85.2	83.6
Swedish	87.5	98.3	92.6	90.5	86.6	88.5	96.6	93.8	95.2	98.6	94.1	96.3

Table 4.7. Normalisation results given in precision (Pre), recall (Rec) and F-score (F).

The baseline case shows the proportion of tokens in the original, historical text that already have a spelling identical to the modern gold standard spelling. In the German evaluation corpus, 84.4% of the historical tokens already have a modern spelling, with a character error rate of 0.16. In the Hungarian corpus on the other hand, only 17.1% of the historical tokens have a modern spelling, with a character error rate of 0.85. At a first glance, the historical spelling in the Hungarian corpus appears to be very similar to the modern spelling, despite the rather high character error rate. A closer look however reveals recurrent differences involving single letter substitutions and/or the use of accents, as for *fiayval* → *fiaval* (fi-a-i-val, son-POSS-1PL-INS, ‘with his/her sons’), *mèghalanac* → *meghalának* (meghal-á-nak, die-PST-3PL, ‘they died’), and *hazaba* → *házába* (ház-á-ba, house-POSS-ILL, ‘to his/her house’).²

²Glossing by the Leipzig Glossing Rules [Lehmann, 1982]. POSS = possessive form, 1PL = first person plural, INS = instrumental case, PST = past tense, 3PL = third person plural, ILL = illative case.

Similarly, the Icelandic corpus also shows a relatively low number of historical tokens with a spelling identical to the modern spelling. Even though the Hungarian and Icelandic texts are older than the English, German, and Swedish texts, the low proportion of tokens with a modern spelling in the Icelandic corpus is rather surprising, since the Icelandic language is generally seen as conservative in spelling. A closer inspection of the Icelandic corpus reveals the same kind of subtle single letter divergences and differences in the use of accents as for Hungarian, as in *ad* → *að* ('to') and *hun* → *hún* ('she').

The simplistic memory approach, relying solely on previously seen tokens in the training data, captures frequently occurring word forms and works surprisingly well, improving normalisation accuracy by up to 63 percentage points. The Levenshtein-based approach in its basic version (referred to as *Lev*), without a normalisation memory, also improves normalisation accuracy as compared to the baseline. However, for all languages, the simplistic memory approach yields significantly higher normalisation accuracy than the more sophisticated Levenshtein-based approach does. This could be partly explained by the fact that frequently occurring word forms have a high chance of being captured by the memory approach, whereas the Levenshtein-based approach runs the risk of consistently normalising high-frequent word forms incorrectly. For example, in the English Levenshtein normalisation process, the high-frequent word form *stonde* has consistently been normalised to *stone* instead of *stand*, due to the larger edit distance between *stonde* and *stand*. The even more common word form *ben*, which should optimally be normalised to *been*, has consistently been left unchanged as *ben*, since the BNC corpus, which is used for dictionary lookup in the English setup, contains the proper name *Ben*. The issue of proper names would not be a problem if a modern dictionary were used for Levenshtein comparisons instead of a corpus, or if casing was taken into account in the Levenshtein comparisons. There would however still be cases left like *stonde* being incorrectly normalised to *stone* as described above, which would be disadvantageous to the Levenshtein-based method. The low recall figures, especially for Hungarian, also indicate that there may be old word forms that are not present in modern dictionaries and thus are out of reach for the Levenshtein-based method, as for the previously discussed Hungarian word form *meghalának*.

When adding a normalisation memory to the Levenshtein-based approach (*Lev+ memory*), the memory is used as a first step in the normalisation process. Only tokens that could not be matched by looking up word forms in the training corpus are normalised by Levenshtein comparisons. The idea is that combining these two techniques would perform better than one approach only, since high-frequent word forms are often normalised correctly through the memory, whereas previously unseen tokens may be handled by Levenshtein comparisons. This combination does indeed perform better for all languages, and for Icelandic this is by far the most successful normalisation method of all.

For the SMT-based approach, it is interesting to note that the simple unigram models in many cases perform better than the more informative bigram models. I also tried adding a normalisation memory to the SMT approach, so that only tokens that are not found in the training corpus are considered for normalisation by the SMT model. This did however not have a positive effect on normalisation accuracy, probably because the training data has already been taken care of by the SMT model, so adding the normalisation memory only leads to redundant information. For four out of five languages, the GIZA++ unigram setting yields the highest normalisation accuracy of all SMT models evaluated. For Hungarian, the GIZA++ bigram model performs marginally better than the unigram model.

The evaluation results presented so far measure normalisation accuracy at a token level, that is the proportion of correctly normalised word forms in a text. These results do however not reflect the proportion of correctly normalised *unique* word forms in the text. Hence, a rather high token-level accuracy could potentially be achieved for a normalisation approach that is able to capture the most frequently occurring word forms in a text, even in cases where a considerable amount of the content words have been incorrectly normalised (or left unnormalised). In particular, it could be assumed that the memory-based approach would benefit from the token-based evaluation setting, since the memory is expected to cover most of the frequently occurring function words, whereas more rare content words may be represented to a less extent. To explore this assumption in more depth, I therefore tried a type-based evaluation as well, measuring the proportion of correctly normalised unique word forms (types) in the evaluation corpus. The results are presented in Table 4.8.

	English	German	Hungarian	Icelandic	Swedish
baseline	44.7	69.2	6.0	35.9	43.2
Lev	60.9	74.8	17.1	65.3	66.1
memory	76.2	86.2	54.1	70.4	59.1
Lev+memory	80.7	87.5	56.9	78.7	73.5
SMT	83.9	92.1	71.2	64.1	80.5

Table 4.8. Normalisation results given in normalisation accuracy at a type level.

As expected, the type-based results are generally lower than the token-based results, indicating that less frequent word forms are harder to handle in the normalisation process. In terms of the relation between the results for different normalisation approaches, the type-based results are however not very different from the token-based results. In both evaluation settings, the SMT-based normalisation approach outperforms the other methods for all languages except for Icelandic, where the Levenshtein-based approach combined with a memory yields the highest normalisation accuracy. Furthermore, the memory-based approach still leads to better results than the Levenshtein-based approach, also in the type-based evaluation setting. This is true for all lan-

guages except for Swedish, where the Levenshtein-based approach surpasses the memory-based approach in the type-based evaluation. One explanation for the fact that the Levenshtein-based approach is more successful for Swedish could be that in the Swedish setting, a morphological dictionary is used for Levenshtein comparisons, meaning that all inflectional forms of a particular word are covered in the lexical resource. For the other languages, Levenshtein comparisons are instead made towards modern language corpora, where only a subset of the possible inflectional forms for a certain word would typically occur in the corpus. This could be problematic, especially in the context of historical text, since some morphological forms may be more frequent in old text than in modern language, due to language change.

Another interesting feature to explore in the evaluation, is the relation between training and test data. In the results presented so far, training and test sets have been extracted from the same source texts. Even though different parts of the texts have been used for training and evaluation, this could potentially affect the evaluation results, since rare word forms and spelling variants used by a specific writer in a specific text have a higher chance of being present in both the training and the evaluation set, than if a totally different text was to be used for evaluation.

To explore how well the different normalisation methods perform on evaluation data extracted from other sources, I evaluated each normalisation approach on a previously unseen text, *Skellefteå Höstting 1771*. This text is also a Swedish court records text, but written in a different part of Sweden in the year 1771. Token- and type-based normalisation results for this test text are given in Table 4.9 (where the gold standard normalisations were manually defined, following the same principles as presented in Chapter 3.2).

	Token-based	Type-based
baseline	74.9	61.3
Lev	84.6	73.6
memory	88.2	68.8
Lev+memory	90.9	76.5
SMT	92.6	84.5

Table 4.9. Normalisation results for out of domain data, given in normalisation accuracy at a token level and at a type level.

These results are not directly comparable to the previously presented results for Swedish, since the new evaluation text is slightly younger than most of the texts sampled in the original evaluation corpus. This means that the proportion of types and tokens with a spelling identical to the modern spelling in the original (unnormalised) text is higher in the new evaluation text. Still, the relation between the results for different normalisation methods is similar to the relation observed for the original evaluation corpus. The SMT-based method still yields the highest results, both in the token-based and in the type-based evalu-

ation setting. Likewise, the memory-based approach yields better results than the Levenshtein-based approach in the token-based evaluation setting, whereas the opposite is true for the type-based setting. For future work, it would be interesting to perform the same kind of evaluation on a new test corpus for the other languages as well. One hypothesis is then that the Levenshtein-based approach may outperform the memory-based approach in terms of normalisation accuracy also for these languages, at least in the type-based evaluation setting, due to a larger proportion of unknown words in the previously unseen test text.

From the results presented throughout this chapter, it is not obvious which normalisation approach to choose for a new language. For Icelandic, the Levenshtein-based approach combined with a normalisation memory leads to the highest normalisation accuracy. For the rest of the languages, the SMT-based approach with the GIZA++ unigram or bigram setting gives the best results. Generally, the Levenshtein-based method could be used for languages lacking access to annotated historical corpora with information on both original and modernised spelling. If, on the other hand, such data is available, the memory-based approach, or the combination of a normalisation memory and Levenshtein calculations, would be likely to improve normalisation accuracy. Moreover, the effort of training a character-based SMT system for normalisation would be likely to further improve the results.

5. Part I: Summary and Conclusion

In the first part of this thesis, I have evaluated four different approaches to spelling normalisation of historical text: a rule-based approach, a Levenshtein-based approach, a memory-based approach, and an SMT-based approach.

The rule-based approach was developed and evaluated for Swedish only. Even though this approach is language-dependent and requires manual efforts for defining the rule set, it has the advantage that no parallel historical-modern corpus is required for rule development, provided that there is some information available on what spelling differences the rules should cover, for instance known spelling reforms and/or historical texts in their original spelling only. I also showed that a relatively small set of hand-crafted normalisation rules based on one single 17th century court records text, had a large positive impact on normalisation results also for texts from other time periods and genres. The choice of text used for developing the rules could possibly be important for success, since older texts generally contain a higher number of instances of differently spelled words. If a too modern text is chosen, the rules generated may thus not be very useful for older texts. For the corpus used in my experiments however, it was shown that the same kinds of spelling variations occurred in texts from all centuries and genres (16th, 17th and 18th century court records and Church documents), only with a slightly different frequency distribution.

The other three normalisation approaches are all language-independent, and were evaluated for five different languages: English, German, Hungarian, Icelandic, and Swedish. The normalisation results varied for the different languages, meaning that it is not entirely clear which approach to choose for a new language, even though the SMT-based approach generally performed the best. A potential problem with the SMT-based approach is that in traditional SMT systems, large amounts of training data is usually required to obtain satisfactory results. Fortunately, in the case of spelling normalisation, translation is performed at a character level, to capture changes in spelling rather than the full translation of words and phrases. Since the set of possible characters and character combinations in a language is far more limited than the set of possible words and word combinations, less training data is needed for this approach. Furthermore, the number of characters in a text is larger than the number of words, meaning that more training data is obtained from the same text when performing character-based translation as compared to word-based translation. In my experiments, I showed that it is possible to achieve good normalisation results with only a small amount of training data, down to 1,000

tokens only. The results also revealed that standard sentence alignment methods may successfully be used for automatically creating training data, in cases where historical-modern parallel data is available in an unaligned form only.

The Levenshtein-based approach to spelling normalisation builds on string similarity calculations between the original historical word form and word forms available in a modern dictionary. This method has the advantage that no historical training data is required, since a modern, monolingual dictionary is used for edit distance comparisons. If a corpus of parallel historical-modern data is available though, this could be used for improving normalisation performance by 1) the inclusion of a normalisation memory, 2) the use of weighted Levenshtein comparisons, and 3) the use of a tuning corpus for setting a threshold value for which maximum edit distance to allow between the original word form and the normalisation candidate(s) extracted from the dictionary. In my experiments, the Levenshtein-based normalisation, combined with a memory, proved to be the most successful normalisation method for Icelandic. It is also interesting to note that the simplistic memory-based approach alone works better than could be expected, yielding higher normalisation accuracy for all languages than the more sophisticated Levenshtein-based approach does. In contrast to the Levenshtein-based method in its basic setting, the memory-based approach does however require access to parallel historical-modern data.

In conclusion, all the proposed approaches to spelling normalisation are successful in increasing the proportion of tokens in the historical text with a spelling identical to the modernised gold standard spelling, thus having the potential of enabling the use of modern NLP tools for analysing historical texts. The rule-based approach proved to be surprisingly useful for texts from other time periods and genres than the text used for rule development. The other approaches were all shown to be language-independent and generally applicable even in cases where only small amounts of training data are available.

Part II:
Linguistic Analysis of Historical Text

6. Verb Phrase Extraction

The primary goal expressed in this thesis, is to present a generic workflow for information extraction from historical text. As described in Section 1.2, the actual information extraction phase in this workflow is based on automatic linguistic analysis of the historical input text, using state-of-the-art NLP tools developed for the modern language. To enable the use of contemporary NLP tools for analysing historical input data, the text is automatically normalised to a more modern spelling, before the NLP tools are applied. In the first part of this thesis, I presented and evaluated four different techniques for converting historical text to a modern spelling. In the second part of the thesis, I focus on the subsequent linguistic analysis step.

The information extraction pipeline is designed to be language-independent and applicable to different information needs. As a proof-of-concept, the pipeline is however applied to the specific information need expressed within the Gender and Work project, where historians are interested in extracting verb phrases describing working activities from Early Modern Swedish text (see further Chapter 1). Thus, the linguistic analysis step is here focused on identifying verb phrases in Swedish text from this time period. This is done in two separate steps: *verb identification* and *complement extraction*. In the verb identification step, all word forms analysed as verbs by the tagger are extracted. In the complement extraction step, complements are assigned to the extracted verb forms, based on the annotation labels given by the parser. The whole verb phrase extraction process is illustrated in Figure 6.1.

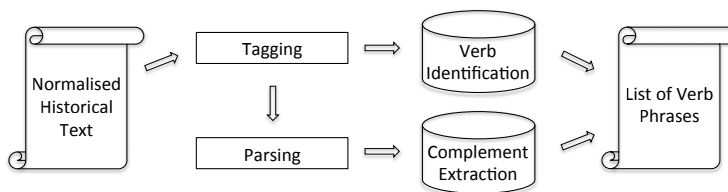


Figure 6.1. Verb phrase extraction from normalised historical input data.

In these experiments, tagging and parsing tools are applied to automatically normalised input data, using the best-performing normalisation approach in accordance with the results presented in Chapter 4, that is the SMT-based method using GIZA++ with unigram models for character alignment. Since the most favourable normalisation method is thus chosen based on the results

achieved in the spelling normalisation experiments, the evaluation corpus used for those experiments may be regarded as inappropriate for further evaluation purposes. To avoid using previously seen data for evaluation, I therefore compiled a new evaluation corpus for the verb phrase extraction experiments, by collecting another 20 random sentences from each subtext of the Gender and Work corpus, in total 300 sentences (non-overlapping with the sentences in the previously used subsets). For training, the previously defined training and tuning sets were merged into a single training corpus, whereas the previously defined evaluation corpus is used for tuning in the verb phrase extraction experiments. The resulting data sets (all extracted from the Gender and Work corpus), and their use in the different experiments, are illustrated in Table 6.1.

Data Set	Tokens	Spelling Normalisation	VP Extraction
A	28,237	Training	Training
B	2,590	Tuning	
C	33,544	Evaluation	Tuning
D	14,672	-	Evaluation

Table 6.1. *Data sets used for training, tuning and evaluation in the spelling normalisation experiments and in the verb phrase extraction experiments, respectively. VP = verb phrase.*

The Gender and Work corpus is not fully linguistically annotated, which would enable a thorough evaluation of tagging and parsing performance. Due to the overall aim of extracting verb phrases from these texts, I have however manually annotated all the verbs and their complements in the tuning and evaluation corpora.¹ This annotation is sufficient for adequately evaluating both verb identification and verb phrase extraction. Verb identification is evaluated in terms of precision and recall measures, by comparing the instances annotated as verbs in the gold standard to the instances annotated as verbs by the tagger. Similarly, verb phrase extraction results are evaluated by comparing the phrases annotated as verbal complements in the gold standard to the annotation labels given by the parser. In both cases, the tagger and the parser are applied to the historical text in its automatically modernised spelling. The results are also compared to the baseline case, in which the tagging and parsing tools are applied directly to the original, unnormalised version of the text. The upper bound is defined as the results achieved when applying the NLP tools to the manually normalised gold standard spelling of the text. In addition, the results are compared to verb phrase extraction results on modern language data.

¹In the ideal case, it would have been desirable to have several annotators annotating the same text, enabling the calculation of inter-annotator agreement to ensure consistent and high quality annotations. Due to limited resources, this was however not possible to accomplish in this study.

6.1 Verb Identification

For the information extraction task at hand, that is automatic extraction of verb phrases describing work from Early Modern Swedish text, the most crucial step is to identify the actual verbs in the text. For verb identification, I perform part-of-speech tagging using HunPOS [Halácsy et al., 2007], a free and open source reimplement of the TnT-tagger developed by Brants [2000]. Both taggers are based on Hidden Markov Models, with trigram language models and suffix guessing algorithms for unknown word forms. In my experiments, I use the HunPOS tagger with a pre-trained language model based on the Stockholm-Umeå Corpus (SUC, version 2.0) [Ejerhed and Källgren, 1997]. Megyesi [2009] has shown that the HunPOS tagger trained on this SUC model is one of the best-performing taggers for (contemporary) standard Swedish text.

As stated in the previous chapter, all the verbs have been manually annotated as such in the evaluation corpus. Thus, the verb identification task may be evaluated based on precision and recall measures, by comparing the instances annotated as verbs in the gold standard to the instances annotated as verbs by the tagger. Precision, recall and F-score values are then calculated by the following formulae:

true positives (tp) = word forms annotated as verb, both in the gold standard and by the tagger

false positives (fp) = word forms annotated as verb by the tagger, but not in the gold standard

false negatives (fn) = word forms annotated as verb in the gold standard, but not by the tagger

$$\text{precision} = \frac{tp}{tp+fp}$$

$$\text{recall} = \frac{tp}{tp+fn}$$

$$\text{F-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

The verb identification results based on HunPOS tagging are presented in Table 6.2. The results show major improvements on the NLP performance if the text is normalised before the tagger is applied. Without normalisation, an F-score of 68.9% is achieved for the verb identification task, as compared to 85.5% for the automatically normalised text. The upper bound results for verb identification on perfectly normalised input data, yield a slightly higher F-score of 90.2%.

The largest improvement for the normalised versions of the text as compared to the unnormalised version is achieved for recall. This is not surprising, as the original text naturally contains a larger proportion of unseen word forms, due to the unfamiliar spelling. Since unseen word forms are gener-

	Precision	Recall	F-score
Unnormalised (baseline)	75.5%	63.3%	68.9%
Automatically Normalised	83.6%	87.5%	85.5%
Manually Normalised (upper bound)	88.3%	92.3%	90.2%
Contemporary Standard Swedish (SUC)	99.1%	99.1%	99.1%
Contemporary Blog Text (SIC)	98.2%	97.6%	97.9%

Table 6.2. *Precision and recall measures for verb identification in Early Modern Swedish text, based on tagging. Baseline = historical text in its original, unnormalised spelling. Automatically Normalised = historical text automatically normalised by the SMT-based approach using GIZA++ with unigram models for character alignment. Upper bound = historical text in its manually normalised spelling.*

ally more likely to be tagged as nouns than verbs, this would result in lower recall for verb identification, whereas verb forms with a spelling identical to the modern spelling would often be recognised as verbs and thus correctly annotated, contributing to keep the precision score at a higher level.

For comparison with contemporary text, I also ran the verb identification experiments on a subset of the SUC corpus, containing those segments in SUC that have been syntactically annotated and manually revised in the Swedish Treebank. In total, this subset includes approximately 20,000 tokens. Since the tagger used in the experiments on verb identification in historical text is trained on the whole of SUC, and I want to avoid evaluating on the same data as the tagger has been trained, I trained a new language model for the tagger in the experiments on contemporary Swedish. In this setting, the training data for the tagger includes all tokens in SUC except for the ~20,000 tokens that are used for evaluation. The results show that verb identification in contemporary standard Swedish is a fairly trivial task for the HunPOS tagger, yielding precision and recall scores around 99.1%. The very high verb identification scores for this evaluation corpus may however be partly explained by the fact that the text contained in the evaluation part of the SUC corpus is sampled from the same source texts as the text contained in the rest of the SUC corpus, meaning that the two data sets are expected to be rather homogenous. Therefore, I also performed experiments on verb identification in the *Stockholm Internet Corpus* (SIC) of blog texts containing in total 13,562 tokens.² Indeed, the verb identification results for the SIC corpus are not as high as for the SUC corpus, but fairly close. Even though the numbers are somewhat lower for historical text, the results are still promising as a basis for further linguistic analysis.

²www.ling.su.se/sic, downloaded 7 October 2015

6.2 Complement Extraction

In the complement extraction phase, the word forms identified as verbs by the tagger are to be assigned the proper complements to form the full verb phrase concealed in the sentence at hand. This is done based on the annotation labels given by a parser. For this purpose, I use MaltParser version 1.6, a data-driven dependency parser developed by Nivre et al. [2006a]. In dependency parsing, the syntactic structure is represented as a directed graph, where the nodes represent lexical items and the arcs represent the dependency relation between the head and its dependent. The arcs are typically labelled, meaning that each arc has a label describing the grammatical function that holds between the head and its dependent. An example borrowed from Nivre [2008] is given in Figure 6.2, illustrating a dependency graph from the Penn Treebank for the sentence *Economic news had little effect on financial markets*. In this sentence, the verb *had* is the root node, with three dependents: *news* is a dependent with the subject relation (SBJ), *effect* is a dependent with the object relation (OBJ), and the full stop is a dependent with the punctuation relation (P). Similarly, the word form *news* has in turn the nominal modifier (NMOD) *Economic* as a dependent, etc.

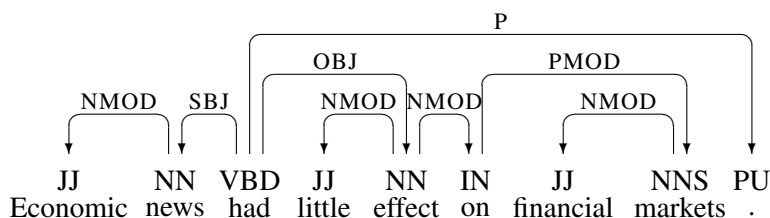


Figure 6.2. Dependency graph borrowed from Nivre [2008], illustrating a dependency graph from the Penn Treebank for the sentence *Economic news had little effect on financial markets*.

MaltParser is typically trained on syntactically annotated corpora, *treebanks*, to induce a parsing model for a particular language or domain. In my experiments, I run the parser with a pre-trained model for parsing contemporary Swedish standard text, based on the Talbanken section of the Swedish Treebank [Nivre et al., 2006b].³ For the verb phrase extraction task, every word form analysed as a verb by the tagger is treated as the head of a verb phrase, and every phrase that the parser has analysed as a dependent of the verb in question is treated as a complement, provided that the dependency relation is labelled with a relevant grammatical function. The following grammatical functions are defined to be relevant in this context (subjects are included only

³The pre-trained model is available at http://maltparser.org/mco/swedish_parser/swemalt.html

if the verb has been analysed as a passive verb by the tagger, in which case the subject is likely to correspond to the direct object in the active voice):

1. Subject (SS) - passive voice only
2. Direct object (OO)
3. Indirect object (IO)
4. Predicative complement (SP)
5. Prepositional complement (OA)
6. Infinitive complement of object (VO)
7. Verb particle (PL)
8. Reflexives (REFL)

Evaluation is performed in terms of precision, recall and F-score based on the extracted verb phrase, as compared to the manually defined annotations given in the evaluation corpus. An authentic example of a (somewhat shortened) manually annotated segment in the corpus is given below:

Nils Jonsson **hade**/**VB**₁ **stullit**/**VB**₂ [OA_{VB2} *honom ifrån* OA_{VB2}] i hans quarn huus [OO_{VB2} 8 *skepper miöll* OO_{VB2}], men tiuffen **bekenne**/**VB**₃, [OO_{VB3} *att han* **togh**/**VB**₄ [OO_{VB4} 4 *skepper* OO_{VB4}] OO_{VB3}]

Nils Jonsson **had**/**VB**₁ **stolen**/**VB**₂ [OA_{VB2} *him from* OA_{VB2}] in his mill house [OO_{VB2} 8 *bushels flour* OO_{VB2}], but the_thief **confesses**/**VB**₃, [OO_{VB3} *that he* **took**/**VB**₄ [OO_{VB4} 4 *bushels* OO_{VB4}] OO_{VB3}]

This sentence has four verbs: *hade* ('had'), *stullit* ('stolen'), *bekenne* ('confesses'), and *togh* ('took'). The first verb, *hade* ('had'), has no complements associated with it in this particular sentence. The second verb on the other hand, *stullit* ('stolen'), has been assigned two complements: the prepositional complement *honom ifrån* ('him from'),⁴ and the direct object *8 skepper miöll* ('8 bushels flour'). The third verb, *bekenne* ('confesses'), has been assigned a complete subordinate clause as a direct object: *att han togh 4 skepper* ('that he took 4 bushels'). Finally, the fourth verb, *togh* ('took'), has been assigned the noun phrase *4 skepper* ('4 bushels') as a direct object. Hence, the gold standard annotation for this sentence suggests four different verb phrases:

⁴For convenience, all adpositional phrases are referred to as *prepositional*, even in cases where the adposition succeeds the noun phrase and would thus normally be defined as a postposition. This is motivated by the fact that postpositional phrases in the historical language variant normally correspond to prepositional phrases in contemporary Swedish.

1	<i>hade</i>	‘had’
2	<i>stullit honom ifrån 8 skepper miöl</i>	‘stolen him from 8 bushels of flour’
3	<i>bekenne att han togh 4 skepper</i>	‘confesses that he took 4 bushels’
4	<i>togh 4 skepper</i>	‘took 4 bushels’

For evaluation of the complement extraction task, I am interested in the ability of the system to correctly extract complements to the word forms identified as verbs in the previous verb identification step. Thus, only those instances where a verb has been correctly identified by the system are considered at this stage, disregarding cases where the system has analysed a word form as a verb, but the human has not, or vice versa. True positives are then defined as cases where there is a non-empty overlap between the automatically extracted verb phrase and the gold standard verb phrase, including cases where both the automatic system and the human annotator have assigned no complements at all to the verb in question. Accordingly, false positives are cases where the system has assigned one or more complements to a verb, but the human has regarded the verb as intransitive or has assigned complements that do not overlap with the ones extracted by the system. False negatives are then cases where the human annotator has assigned one or more complements to a verb, whereas the system has not. The results are presented in Table 6.3.

	Precision	Recall	F-score
Unnormalised (baseline)	71.9%	36.1%	48.1%
Automatically Normalised	76.3%	50.5%	60.7%
Manually Normalised (upper bound)	77.9%	55.4%	64.8%
Contemporary Standard Swedish (SUC)	91.6%	77.5%	84.0%

Table 6.3. *Precision and recall for verb phrase extraction based on parser output. Baseline = historical text in its original, unnormalised spelling. Automatically Normalised = historical text automatically normalised by the SMT-based approach using GIZA++ with unigram models for character alignment. Upper bound = historical text in its manually normalised spelling.*

The results show a substantial improvement for the verb phrase extraction task, when normalisation is performed prior to tagging and parsing. This is especially true for recall, where more verb forms are correctly identified when the spelling is more similar to the modern spelling. Furthermore, the results for automatically normalised input data are close to the upper bound results achieved for the gold standard spelling. This indicates that the spelling normalisation step has been successful in transforming the original word forms into forms that are recognisable by the contemporary NLP tools, and that other measures are called for in order to obtain evaluation scores that are closer to the scores achieved for the contemporary Swedish SUC corpus.

7. Valencies for Improved Verb Phrase Extraction

As shown throughout the second part of this thesis, spelling normalisation has a positive effect on both tagging and parsing, when NLP tools developed for the modern language are used for analysing historical text. However, the verb phrase extraction results presented for Swedish reveal that the parser still has problems extracting the correct complements associated with a verb in a sentence. This is true even when the parser is applied to manually normalised data, with a recall of approximately 55.4% for the verb phrase extraction task. This indicates that other characteristics of historical text, apart from spelling, could be problematic for the parser to handle, for instance word order differences and significantly longer sentences in the Early Modern Swedish corpus than in present-day Swedish. The latter issue is further aggravated by sentence segmentation problems due to inconsistent use of punctuation (as discussed further in Section 2.1.6).

One could think of several ways to deal with these problems, such as re-ordering methods for converting the word order into a more standard one, or sentence splitting techniques for dividing the original sentence into shorter subsentences. In this section I propose a method aiming at improving verb phrase extraction results by adding verb valency information to the extraction process, as illustrated in Figure 7.1.

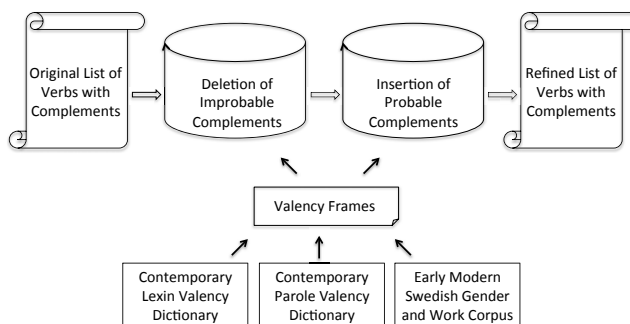


Figure 7.1. Valency-based post-processing step to verb phrase extraction from historical text.

In the valency-based verb phrase extraction approach, verb valency information is provided in a post-processing step, where verbal complements extracted

by the parser are removed if not consistent with the valency frames associated with the verb in question. Likewise, verbal complements are added to the extracted phrase, based on the valency frame of the verb combined with specific types of phrases found in the near context of the verb in the sentence. In short, improbable complements suggested by the parser are removed from the extracted phrase, whereas probable complements not found by the parser are added to the extracted phrase.

7.1 Data

The data sets used in the valency-based verb phrase extraction experiments are presented in Table 7.1.

Verb Valency Dictionaries	
	Verb Entries
Lexin	3,281
Parole	4,304
Historical Corpus Data	
	Tokens
GaW Training	30,827
GaW Tuning	33,544
GaW Evaluation	14,672

Table 7.1. Data sets used in the valency-based experiments.

Verb valency frames are extracted from three sources: the contemporary Lexin valency dictionary,¹ the contemporary Parole valency dictionary,² and the training part of the Early Modern Gender and Work corpus (GaW). The GaW corpus is added to the more conventional contemporary valency dictionaries to improve coverage. Some word forms in historical text are not frequent enough in contemporary language to occur in modern dictionaries. Examples from the GaW corpus are *absentera* (old word for ‘be absent’), *umgälla* (old word for ‘suffer for’), and *ärna* (old word for ‘intend to’). Moreover, the meaning of verbs tend to change over time, and it is not obvious that verb valency frames for present-day Swedish also hold for historical Swedish. An example from the GaW corpus is the verb *slå* (‘hit’) which in both Lexin and Parole is listed as a monotransitive verb (‘to hit someone’). In the GaW corpus however, it is repeatedly used as a ditransitive verb, as in *Sedhan hadhe Erich OluffSon slaghit Pelle Pederssonn tre blånader* (‘Then Erich OluffSon had hit Pelle Pederssonn three bruises’). A comparison between the valency frames present in the GaW corpus and the frames present in the Lexin dictionary shows that only

¹http://spraakbanken.gu.se/lexin/valens_lexikon.html

²<http://spraakbanken.gu.se/swe/resurs/parole>

16% of the verb forms that are present in both the old and the modern resource (108 out of 675 verb forms) have equal valency frames. If instead comparing the two contemporary valency dictionaries to each other, approximately 89% of the valency frames are equally defined in the two resources.

There are no valency frames explicitly stated in the GaW corpus. Instead, valency information is extracted from the corpus based on the complements that have been manually assigned to the verb forms occurring in the corpus. As previously stated, all the verbs in the three subcorpora have been manually annotated as such, and all complements adhering to the verbs are annotated with labels denoting subject (for passive verbs only), direct object, indirect object, prepositional complement, infinitive complement, subject predicative, verb particle, and reflexive pronoun. For each verb form occurring in the training part of the corpus, all the complement types that have been assigned to this particular verb form in any sentence are stored as possible complements to the verb in question. Similar to the valency information provided in the contemporary dictionaries, the specific preposition or particle that could be seen as the head of the prepositional complement or particle construction is also stored in the valency frame. For example, the verb *slå* ('hit') has two different types of particle complements in the valency frame extracted from the GaW corpus, one with the particle *igen* and one with the particle *ihjäl*, as illustrated below:

PL-igen *H:r Carl slogh dörren igen* ('Mr Carl **shut** the door **closed**')
PL-ihjäl *Han slogh hans fader ihjell* ('He **beat** his father **to death**')

Both in the Lexin dictionary and in the Parole dictionary, verb valency frames are connected to the present tense form of the verb only, without information on other inflectional forms of the verb. In the verb phrase extraction process however, there is a need to connect whatever inflectional form of the verb that is used in a sentence to the correct valency frame. For broader coverage of the valency dictionaries, I therefore expand the present tense forms to other inflectional forms, based on the Saldo dictionary of present-day Swedish word forms (version 2.0) [Borin et al., 2008] and the SUC corpus of contemporary Swedish (version 2.0) [Ejerhed and Källgren, 1997]. By comparing the present tense verb form given in Lexin or Parole to the morphological and inflectional information present in the Saldo dictionary, it is possible to extract a lemma corresponding to the verb form, and from that lemma all the inflectional forms adhering to that lemma. For verb forms not found in the Saldo dictionary, the SUC corpus is consulted. Since this corpus is annotated with lemma information, it is possible to group together all the inflectional forms occurring in the corpus that adhere to the same lemma. For Lexin and Parole verb forms neither found in Saldo nor SUC, only the present tense form of the verb is stored with its corresponding valency frame.

For the GaW corpus, there is a similar problem in that only those verb forms that occur in the corpus will be assigned a valency frame, and if several

forms of the same verb occur in the corpus, these will be assigned valency frames separate from each other. To deal with this, I use the same method of comparison to Saldo and SUC for retrieving the full set of word forms associated with a verb form, assigning the same valency frame to all verb forms belonging to the same lemma. In this process, the manually normalised form of each verb is used for comparison towards Saldo and SUC, to avoid mismatches due to spelling variation in the historical corpus.

It could be argued that instead of generating all full forms for a verb, it would be more efficient to perform lemmatisation prior to comparison. This would however potentially impose more ambiguity to the valency frames, since word forms in the SUC corpus are associated with their base form rather than the actual lemma. This means that present tense forms such as *är* ('is') and *varar* ('lasts') are both associated with the same base form *vara* ('to be/to last'), even though their inflectional paradigms and valency frames differ significantly. For properly lemmatised sources, these word forms would instead have been associated with different lemmas, e.g. *vara1* and *vara2*.

Table 7.2 lists the total number of entries in the languages resources used in the valency-based experiments, before and after full form expansion. It also shows the number of verb forms found in Saldo and SUC respectively, during the process of expanding the valency frames to more inflectional forms. Note that SUC is only consulted for verbs not found in Saldo, which explains the lower numbers presented for SUC. Since the test set of the GaW corpus will only be used for evaluation, no expansion to inflectional forms is needed for this particular data set.

	Verbs	In Saldo	In SUC	Not found	Expanded Forms
Lexin Dictionary	3,281	3,181	33	67	42,545
Parole Dictionary	4,304	4,263	26	15	32,640
GaW Training	1,329	1,168	14	147	10,032
GaW Tuning	1,410	1,245	15	150	10,394
GaW Evaluation	987	n/a	n/a	n/a	n/a

Table 7.2. Number of verb forms in the language resources, before and after full form expansion. *Verbs* = Number of distinct verb forms in the original valency resources. *In Saldo* = Verb forms found in Saldo during the process of expanding the valency frames to more inflectional forms. *In SUC* = Verb forms found in SUC. *Not found* = Verb forms found neither in Saldo nor SUC. *Expanded Forms* = Number of distinct verb forms after full form expansion.

In the following experiments, I use the training part of the corpus as a basis for valency frames during model selection, where the tuning part is used for repeated testing. In the final evaluation, the training and tuning sets are merged to a combined valency resource, and the evaluation part of the corpus is used for testing.

7.2 Deletion of Improbable Complements

As previously discussed, certain characteristics of historical text make it difficult for the parser to correctly extract the complements of a verb. One way to improve precision in the complement extraction phase would be to automatically filter away extracted complements that do not conform to the valency frame of the verb in question. I perform deletion experiments for all complement types extracted from the parser except for subjects, since a verb is typically expected to have a subject. I try the following five deletion settings:

1. **Lexin**

For each extracted complement, if the head verb is present in the Lexin valency dictionary and the valency frame in Lexin does not allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

2. **Parole**

For each extracted complement, if the head verb is present in the Parole valency dictionary and the valency frame in Parole does not allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

3. **GaW Corpus**

For each extracted complement, if the head verb is present in the training part of the GaW corpus, and none of the occurrences in the corpus contain a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

4. **All combined**

For each extracted complement, if none of the resources that list the head verb allow for a complement of the type indicated by the parse label, the complement is removed from the extracted verb phrase.

5. **All one-by-one**

For each extracted complement, if the best-performing resource that lists the head verb does not allow for a complement of the type indicated by the parse label, the complement is removed. Performance is here measured in terms of complement extraction F-score in the first three experiments.

7.3 Insertion of Probable Complements

Apart from filtering away unlikely complements extracted by the parser, I also aim at inserting probable complements not found by the parser, by searching the parsed sentence for words and phrases that match the valency frame of the

head verb, but which have not been extracted by the parser. Since the word order is more varying in Early Modern Swedish than in present-day Swedish, all complements are searched for both to the left and to the right of the head verb.

In the insertion experiments, I focus on phrasal verbs in the broader sense, including particles, reflexives, and prepositional complements. The motivation for including these three complement types in the insertion experiments is that these are relatively easy to recognise in a sentence. Furthermore, if for example a reflexive pronoun is found close to the verb in the sentence, and the valency frame suggests a reflexive pronoun, then the probability that this reflexive belongs to the verb is rather high. The same argument holds for prepositional phrases containing the expected preposition to form a prepositional complement, and for prepositions or adverbials identical to a particle expected by the valency frame of the verb. For direct objects on the other hand, even if a noun phrase is found close to the verb, it would still be hard to determine whether this noun phrase actually corresponds to a direct object, since noun phrases may occur with many different functions in a clause, and the word order is not fixed, particularly not for historical text. Therefore one would run a high risk of extracting for example the subject noun phrase instead of the direct object noun phrase, especially for languages like Swedish, where subject/object distinctions are not manifested morphologically other than for pronouns. Furthermore, direct objects are not always expressed in the form of noun phrases, but are quite often expressed as for instance clauses, as in the following example from the Gender and Work corpus: *fordra at Barnet skal döpas hemma* ('demand that the Child should be christened at home'). Similarly, indirect objects and subject predicatives may also be expressed in varying ways, and infinitive complements are often ambiguous to other functions. Thus, these categories are excluded from the insertion experiments.

In accordance with the arguments given above, the following three experiments are performed for insertion of probable complements:

1. Insertion of prepositional complement

If the valency frame of the head verb (in any of the three valency resources) allows for a prepositional complement, and a prepositional phrase containing the expected preposition is found either to the left or to the right of the head verb, this prepositional phrase is added to the extracted verb phrase with a prepositional complement label.

2. Insertion of particle

If the valency frame of the head verb allows for a particle, and a word that is identical to the expected particle and tagged as preposition or adverb is found either to the left or to the right of the head verb, this preposition or adverb is added to the extracted verb phrase with a particle label.

3. Insertion of reflexive pronoun

If the valency frame of the head verb allows for a reflexive pronoun, and the word form *sig* ('oneself'), or the alternative historical spelling *sigh*, is found either to the left or to the right of the head verb, this word form is added to the extracted verb phrase with a reflexive label. It could be noted that this matching procedure captures the reflexive pronoun in the third person only, whereas the first person and second person forms are excluded from the search. This is motivated by the fact that in Swedish, the word form *sig* is unambiguously a reflexive pronoun, whereas the second and third person forms *mig* ('myself') and *dig* ('yourself') are ambiguous word forms. In some cases they are used as reflexive pronouns, as in *jag kliade mig* ('I scratched myself'). They may however also be used as the object form of a personal pronoun, as in *hon kliade mig* ('she scratched me'). To avoid incorrect reflexive pronoun interpretations due to this ambiguity, only the third person form is thus included in the search.

7.4 Model Selection for Deletion

In the model selection phase, I try different strategies for deletion of improbable complements, using the training part of the corpus as a basis for valency frames, and the tuning part of the corpus for testing. In the previous chapter, verb phrase extraction was evaluated in terms of precision, recall, and F-score for the extracted verb phrase as a whole, where true positives are cases where there is a non-empty overlap between the automatically extracted verb phrase and the verb phrase given in the gold standard. In the model selection phase for deletion, I am more interested in looking at the separate complements building up the phrase, with the aim of removing superfluous complements, but still keeping valid complements. Hence, a more fine-grained evaluation metrics is called for in this setting. During the model selection phase, precision, recall and F-score are therefore based on the extracted complements, instead of the extracted phrase as a whole. When comparing the automatically extracted verb phrases to the manually annotated phrases in the gold standard, true positives are defined as correctly extracted complements, that is the automatically extracted phrase is present in the gold standard as well. Likewise, false positives are complements extracted by the system that are not present in the gold standard, whereas false negatives are complements that are present in the gold standard but not extracted by the system. Since I am specifically aiming at extracting the correct complements, intransitive verbs that were also identified as intransitive by the extraction system will not contribute to the set of true positives. Intransitive verbs for which the system has extracted complements will however contribute to the set of false positives, whereas verbs

identified as intransitive by the system though complements are present in the gold standard will add to the set of false negatives.

I also make a distinction between labelled and unlabelled precision and recall. For labelled precision and recall, it is required that the correct label (direct object, prepositional complement etc.) has been assigned to the complement in order for the extracted segment to be regarded as correct. For unlabelled precision and recall on the other hand, a complement is regarded as correctly extracted based on the word sequence alone, regardless of what label the parser has assigned to the complement. The unlabelled results are highly relevant, since the typical end user would be a historian or other researcher in humanities, searching for phrases with a specific semantic content, in which case the underlying syntactic annotation labels are of no interest.

For both labelled and unlabelled results, partial matches are regarded as true positives, since these would normally make the user aware of a relevant phrase, even in cases where the complete phrase has not been detected. True positives thus include the following cases, with authentic examples from the Gender and Work corpus:

- **Exact match**

The extracted phrase is identical to the gold standard phrase.

Gold complement: *2 klimpar smör*

Extracted complement: *2 klimpar smör*

‘2 lumps of butter’

- **Substring**

The extracted phrase is a substring of the gold standard phrase.

Gold complement: *de penningar och medel*

Extracted complement: *medel*

‘(the money and) **resources**’

- **Superstring**

The gold standard phrase is a substring of the extracted phrase.

Gold complement: *detta*

Extracted complement: *detta efter honom*

‘**this** (after him)’

- **Overlap**

There is a non-empty overlap between the extracted phrase and the gold standard phrase.

Gold complement: *förswagat ock förtrygt*

Extracted complement: *nogh förswagat*

‘(probably) **weakened** (and oppressed)’

As previously stated, the purpose of the model selection phase is to try different strategies for deletion of improbable complements, using the training part of the corpus as a basis for valency frames, and the tuning part of the corpus for testing. First, I want to find out which of the five deletion settings listed in Section 7.2 that leads to the best results. I therefore run experiments where deletion is performed for all complement types (except subject), evaluating the results for each setting separately. The results for unlabelled complement extraction are summarised in Table 7.3. Note that the model selection scores given in Table 7.3 are not comparable to the scores given for the original complement extraction method described in Section 6.2, since I take each single complement into account during the model selection phase, rather than evaluating the verb phrase as a whole.

	Precision	Recall	F-score
Baseline	53.30	51.22	52.24
Lexin	59.76	35.44	44.49
Parole	55.22	38.49	45.36
GaW corpus	57.51	46.77	51.59
All combined	56.62	47.76	51.81
All one-by-one	57.64	46.01	51.17

Table 7.3. *Unlabelled model selection results for deletion of improbable complements using different settings, evaluated on the tuning part of the Gender and Work corpus. Baseline = Complement extraction results without the valency-based post-processing step.*

As could be expected, deletion of improbable complements leads to an increase in precision, at the cost of a decrease in recall. Recall varies to a greater extent than precision. For the largest resource, that is the Lexin dictionary, precision is the highest, but recall is very low. This could indicate that a large proportion of verb forms are found in the Lexin dictionary, but with valency frames that do not correspond to the way the verbs are used in the text, meaning that complements are erroneously deleted. This confirms the initial hypothesis that due to language change, contemporary dictionaries are not sufficient for providing valency information. Further arguments for this hypothesis is the fact that even though the training part of the Gender and Work corpus is by far the smallest valency resource, using only this resource for defining verb valency frames results in a substantially higher F-score value than using Lexin or Parole. In fact, the F-score results for using the Gender and Work corpus alone are almost as high as for using all resources combined. Since all methods improve precision as compared to the baseline, I regard the combined method as the best-performing method, since this method has the highest recall and also the highest F-score. But the surprisingly high results

for the Gender and Work corpus suggest that using a historical corpus only could yield comparable results.

In the next round of experiments, I want to explore what complement types should be included in the deletion process. The hypothesis is that some complement types may be more thoroughly covered in the valency resources than others. If so, deletion of complements may only be a successful method for some complement types, whereas others should be left unmodified in the deletion process. To test this hypothesis, I try deletion for each complement type separately, keeping only those that improve F-score as compared to the baseline system. These experiments are run with the combined setting, and the results are presented in Table 7.4. As seen from the table, only deletion of direct objects and subject predicatives are successful in improving the F-score value as compared to the baseline. When only these two categories are included in the deletion process, an F-score of 52.50% is achieved, as compared to the baseline F-score of 52.24%.

	Precision	Recall	F-score
Baseline	53.30	51.22	52.24
A) direct object	54.29	50.62	52.39
B) indirect object	53.37	50.92	52.12
C) prep compl	56.06	47.51	51.43
D) inf compl	53.34	51.02	52.15
E) subj predicative	53.93	50.84	52.34
F) particle	53.45	50.84	52.11
G) reflexive	53.19	50.50	51.81
A + E	54.96	50.25	52.50

Table 7.4. *Unlabelled model selection results for deletion of improbable complements for the setting ‘all combined’, evaluated on the tuning part of the Gender and Work corpus, when varying the complements included for deletion. Baseline = Complement extraction results without the valency-based post-processing step.*

7.5 Model Selection for Insertion

As described in Section 7.3, the insertion experiments are targeted at particles, reflexives, and prepositional objects. Whenever the valency frame of the head verb in the extracted phrase allows for a complement of the specified type, the parsed sentence is searched for words and phrases matching the complement at hand. Insertion is a bit more tricky than deletion, with a high risk of accidentally inserting phrases that happen to match the search criteria, even though the phrase has no syntactic relation to the verb form at hand. This is especially true in historical text, where sentence boundaries and other punctuation marks

are not always as explicit as in modern text. To narrow the search space, I try the following restrictions to the original insertion strategy, based on findings in the tuning part of the Gender and Work corpus:

1. Inclusion of stop verbs, for which no complements are to be added. The set of stop verbs includes all word forms belonging to any of the lemmas *vara* ('be'), *bli* ('become'), *ha* ('have') and *finnas* ('exist'). This stop list was empirically defined by compiling a list of the verb forms that most often lead to incorrect insertion of complements when running the insertion experiments with no restrictions.
2. Prohibiting punctuation to occur between the head verb and the candidate complement.
3. Inclusion of a distance threshold, defining the number of tokens that may come in between the head verb and the candidate complement. I tried a number of different thresholds, out of which a threshold of 5 tokens turned out to yield the best results.

The insertion results are presented in Table 7.5, showing that without any restrictions in the insertion process, recall can be increased from 51.22% to 53.63%, for the tuning part of the Gender and Work corpus. This is however at the expense of a substantial drop in precision from 53.30% to 37.47% as compared to the baseline system.

	Precision	Recall	F-score
Baseline	53.30	51.22	52.24
Unrestricted Insertion	37.47	53.63	44.12
A) Insertion + Stopwords	45.69	53.41	49.25
B) Insertion + Punctuation	47.81	53.04	50.29
C) Insertion + Threshold	51.42	52.54	51.97
Insertion + A + B + C	52.57	52.45	52.51

Table 7.5. *Unlabelled model selection results for insertion of probable complements, evaluated on the tuning part of the Gender and Work corpus. Baseline = Complement extraction results without the valency-based post-processing step.*

Restrictions in the form of A) stopwords for which no complements are inserted, B) prohibition of punctuation between the head verb and the candidate complement, and C) defining a threshold for how many tokens are allowed to occur between the head verb and the candidate complement, all have a positive effect on precision and F-score. Thus, in the best setting, where all three restrictions are implemented, a precision of 52.57% is achieved, with a recall of 52.45%. Also note that the model selection scores given in Table 7.5 are not comparable to the scores given for the original complement extraction method

described in Section 6.2, since I take each single complement into account during the model selection phase, rather than evaluating the verb phrase as a whole.

To find out what complements should be included for insertion, I also try insertion for each complement type separately. As seen from Table 7.6, the best results are achieved when all three complement types are included for insertion.

	Precision	Recall	F-score
Baseline	53.30	51.22	52.24
A) prep compl	52.70	51.86	52.28
B) particle	53.23	51.28	52.24
C) reflexive	53.20	51.75	52.46
A + C	52.60	52.39	52.49
A + B + C	52.57	52.45	52.51

Table 7.6. *Unlabelled model selection results for insertion of probable complements, evaluated on the tuning part of the Gender and Work corpus, when varying the complements included for insertion. Baseline = Complement extraction results without the valency-based post-processing step.*

7.6 Evaluation

Table 7.7 presents the complement extraction results on the evaluation corpus, with the training and tuning part of the Gender and Work corpus merged into a single historical valency resource. Results are presented for the baseline system (without additional deletion or insertion), for the best deletion setting as argued in Section 7.4, for the best insertion setting as argued in Section 7.5, and for both deletion and insertion combined.

As expected, performing only deletion of complements leads to an increase in precision at the expense of a decrease in recall. Performing only insertion on the other hand leads to an increase in recall without decreasing precision, demonstrating that inserting complements introduces true positives to a higher extent than false positives, which is satisfactory. In fact, in the unlabelled case, performing only insertion results in a slightly higher F-score value than the combination of deletion and insertion. However, the best precision is achieved when both deletion and insertion are performed, yielding a precision of 63.04%, as compared to 61.82% for the baseline system. This setting also improves both precision and recall as compared to the baseline.

	Unlabelled		
	Precision	Recall	F-score
Baseline	61.82	48.68	54.47
Deletion	63.02	47.61	54.24
Insertion	61.88	50.80	55.80
Deletion + Insertion	63.04	49.74	55.61
	Labelled		
	Precision	Recall	F-score
Baseline	53.25	38.34	44.58
Deletion	54.75	37.97	44.84
Insertion	53.67	40.45	46.13
Deletion + Insertion	55.12	40.08	46.41

Table 7.7. Precision and recall measures for complement extraction based on parser output. Baseline = Complement extraction results without the valency-based post-processing step.

The results presented throughout this chapter have so far been calculated at a complement level, comparing the automatically extracted complements to the complements given in the manually annotated gold standard. Table 7.8 presents the precision and recall scores for the extraction of the full verb phrase, following the same evaluation metrics as is described in Section 6.2.

	Precision	Recall	F-score
Baseline	76.3%	50.5%	60.7%
Deletion	77.7%	50.7%	61.3%
Insertion	76.6%	51.7%	61.7%
Deletion + Insertion	77.9%	51.9%	62.3%

Table 7.8. Precision and recall measures for verb phrase extraction based on parser output. Baseline = Verb phrase extraction results without the valency-based post-processing step.

Even though these results also show modest improvements, it is interesting to note that performing deletion only or insertion only, improve both precision and recall as compared to the baseline system. The best results are achieved with both deletion and insertion combined, increasing F-score from 60.7% for the baseline system to 62.3% for the refined system.

8. Part II: Summary and Conclusion

In the second part of this thesis, I have presented an approach to linguistic analysis of historical text using taggers and parsers developed for the modern language. The overall aim of performing tagging and parsing in this context is to be able to use the linguistic analysis results as a basis for information extraction from the historical input text. In order to avoid poor results due to spelling variation, the input text is automatically normalised to a more modern spelling, before the NLP tools are applied (using the best-performing normalisation method in accordance with the results presented in the first part of the thesis).

The information extraction pipeline is designed to be applicable to different languages, genres, time periods, and information needs, as long as there are a tagger and a parser available for the modern language variant. In my experiments, I focus on the information need arisen within the Gender and Work project, where historians aim to store Early Modern Swedish phrases describing working activities in a database. Since the historians have concluded that these phrases most often are verb phrases, my experiments for the linguistic analysis step are geared towards automatic extraction of verb phrases from Early Modern Swedish text. Thus, tagging is in this context used for the specific task of verb identification, whereas parsing experiments are conducted for the task of assigning verbal complements to the verbs identified by the tagger.

From the results, it is obvious that normalisation has a large positive effect on both tagging and parsing of historical input data. For verb identification based on tagging, the F-score value increases from 68.9% for unnormalised input data to 85.5% for the automatically normalised version of the input text. For the verb phrase extraction task, where words and phrases that the parser has analysed as dependents of a particular verb are stored as complements to the verb in question, the F-score value increases from 48.1% for unnormalised input data to 60.7% for the automatically normalised version.

Furthermore, for both tagging and parsing, the performance on automatically normalised input data is close to the upper bound results for manually normalised text, but still far from the results for contemporary standard Swedish text. This leads to the conclusion that the automatic spelling normalisation step is satisfactory, and that in order to achieve even better results in the linguistic analysis step, non-spelling-related characteristics of the input text need to be taken into account. In an attempt to further improve the verb phrase extraction results, I therefore experimented on including verb valency information in a post-processing step to the verb phrase extraction workflow.

Based on verb valency frames extracted from historical corpora as well as from contemporary valency dictionaries, I defined a method for automatically deleting verbal complements that are extracted by the parser, if not coherent with the valency frame of the head verb. Likewise, complements that are not found by the parser are added to the extracted phrase, if the valency frame allows for a missing complement and an appropriate phrase is found close to the verb in the sentence at hand. By automatically deleting and inserting complements based on verb valency information, the verb phrase extraction step is improved, increasing the F-score value from 60.7% to 62.3%, as compared to the baseline system where no valency information is used in the extraction process. The experiments also show that the historical corpus has the largest positive effect on the verb phrase extraction results, even though the contemporary dictionaries cover more verb forms. This could be due to several reasons. For example, the valencies extracted from the historical corpus will only contain those complements that are actually used together with a specific verb in the training corpus, thus capturing the most frequent verbal constructions in the language. The contemporary valencies on the other hand, have been extracted from valency dictionaries, where both obligatory and optional complements are listed, regardless of how commonly used they are. Another reason could be that language changes over time, and valency frames for the present-day language may thus not be enough to cover the syntax in historical text.

Part III:
Information Extraction from Historical Text

9. Verb Phrase Ranking

So far, I have presented methods for automatic normalisation of historical text to a more modern spelling, succeeded by linguistic analysis of the normalised text using taggers and parsers developed for the modern language. In the third part of this thesis, I address the task of information extraction from historical text, based on the linguistically analysed version of the text. Similar to the experiments conducted in the second part of the thesis, I focus the information extraction experiments on the problem of finding verb phrases describing work in Early Modern Swedish text (even though the principle of applicability to other languages and information needs is an important aspect for the methods chosen for implementation). As presented in Chapter 1, this particular information need has arisen within the *Gender and Work* project, where historians are storing information in a database on what men and women did for a living in the Early Modern Swedish society (i.e. approximately 1550–1800). During this work they have found that working activities in their source material are most often expressed in the form of verb phrases, such as *to fish herring* or *to sell clothes* [Ågren et al., 2011].

In the ideal case, it would be desirable to extract all verb phrases from a historical text, correctly classify each instance as either describing work or not, and finally present all phrases denoting work, and no other phrases, to the end user. In reality, this is however a tricky task. Even though I do have access to a database of phrases previously extracted by the historians as describing work, this does not guarantee that I automatically know how to categorise similar phrases occurring in other texts. For example, the verb *köpa* (‘to buy’) has sometimes been extracted by the historians as indicating a working activity related to trade, whereas in other contexts, people buy things for non-commercial reasons, in which case the historians have chosen not to store this particular instance of the verb *köpa* in the database. In previously unseen texts, there will also most certainly be previously unseen word forms present, which a classifier would not know how to handle. This problem is further aggravated by the high degree of spelling variation in historical text, combined with inconsistently extracted phrases in the gold standard database. One example of inconsistencies occurring in the database is illustrated by the following two phrases, that are both stored in the database:

haffua lyfftat kitilen aff elden (‘lifted the pot from the fire’)
pighan haffua lyfftat kitilen aff elden (‘the maid lifted the pot from the fire’)

Both excerpts are linked to the same original segment, and annotated as describing the working activity of *lifting a pot*. Inconsistencies like these may be due to different excerptors having separate intuitions on the definition of a phrase, or to the same excerptor making different judgements at different occasions.

To cope with the fuzzy definitions of a phrase in the database, and the risk of running into data sparseness problems, I try a ranking approach to the information extraction task, instead of doing a binary classification into phrases denoting work versus phrases not denoting work. In the ranking approach, I aim to present those verb phrases that *most probably* describe work at the top of the results list, whereas phrases that are less likely to describe work will be presented further down in the list.

Even though the main task here is ranking, the training data available is not ranked but consist of verb phrases classified as describing work (phrases extracted to the database by the historians), and verb phrases that do not describe work (verb phrases present in the source text but not extracted to the database). This binary classification poses special challenges in training the ranking system. I try three different approaches to verb phrase ranking, based on:

1. conditional probability
2. log likelihood ratio
3. bag-of-words classification

All ranking techniques are applied to automatically extracted verb phrases, as described in Chapter 6. Furthermore, word frequencies used in the ranking process are calculated based on the automatically normalised spelling of the phrase, using the best-performing normalisation method for Swedish as defined in Chapter 4, that is the SMT-based method using GIZA++ with unigram models for character alignment.

9.1 Data

For the verb ranking task, I need access to training and evaluation data in the form of positive and negative instances, where positive instances are verb phrases classified as describing work, and negative instances are verb phrases classified as not describing work. For this purpose, the historians have provided a special subset of the Gender and Work data of Swedish court records and church documents from the Early Modern period. This subset consists of text snippets, referred to as *cases* by the historians. Each case typically contains 4–5 sentences, and comprises at least one phrase describing a working activity. The snippets have been manually analysed by the historians, and those phrases that were judged as denoting work are stored in the Gender and

Work database, with information on which case the phrase has been extracted from. This means that I have access both to the source text, and to the phrases within this text that actually describe work. By automatically extracting all the verb phrases from the text snippets using the methods described in the first and second parts of this thesis, it is also possible to infer what verb phrases in the corpus that have not been stored in the database, and thus have been judged **not** to describe work.

Even though it has been stated that working activities in the Gender and Work database are most often expressed in the form of verb phrases, the phrases stored in the database do not always contain a verb. Common non-verb phrases in the database are:

- **noun phrases**
träägårdz dräng på gården
'garden servant at the farm'
- **present participles**
kyrkiotakets reparerande
'repairing of the church roof'
- **past participles**
avlönad vid Gripsholm 1572
'paid at Gripsholm 1572'

Since the verb-oriented approach explicitly aims at extraction of verb phrases, only phrases in which the tagger is able to identify a verb have been included in the data sets, both for training and for evaluation. The resulting data sets used in my experiments are presented in Table 7.1.

	Sentences	Verb Phrases In Total	Positive Instances
Training	10,623	37,606	10,241
Evaluation	1,358	4,770	1,254

Table 9.1. *Data sets used in the verb phrase ranking experiments. Sentences = Total number of sentences in the data set. Verb Phrases In Total = Total number of verb phrases in the data set. Positive Instances = Proportion of verb phrases defined as describing work.*

As seen from the table, approximately 27% of the verb phrases in the corpus are phrases describing work. It should however be noted that this subset of the corpus is biased towards phrases describing work, since the corpus does not comprise the whole source documents, but only those sections within the documents that actually contain some element describing work.

9.2 Conditional Probability

In the *conditional probability* approach, the probability that a verb phrase describes work, given the verbs present in the phrase, is estimated. For every verb in the phrase to be ranked, the probability that this verb describes a working activity is then estimated using the following formula:

A = number of times the specific verb is part of a verb phrase judged as describing work in the training corpus

B = total frequency of the verb in the training corpus

$$P(A|B) = \frac{P(A \cap B)}{B}$$

I try two different ways of ending up with a final ranking score for the phrase. In the first approach, the final ranking score is defined as the conditional probability for the verb with the highest conditional probability score in the phrase. In the second approach, the final ranking score is defined as the average conditional probability score over all the verbs in the phrase. Furthermore, the conditional probability approach is applied both to purely tokenised data (after automatic spelling normalisation), and to lemmatised data. Lemmatisation is performed on the automatically normalised version of the text, using lemma information present in the Saldo dictionary (version 2.0) [Borin et al., 2008], and the SUC corpus (version 2.0) [Ejerhed and Källgren, 1997]. In total, this results in four different settings for the conditional probability approach:

1. using tokenised input data, with the final ranking score defined as the conditional probability for the verb with the highest conditional probability score in the phrase
2. using tokenised and lemmatised input data, with the final ranking score defined as the conditional probability for the verb with the highest conditional probability score in the phrase
3. using tokenised input data, with the final ranking score defined as the average conditional probability score over all the verbs in the phrase
4. using tokenised and lemmatised input data, with the final ranking score defined as the average conditional probability score over all the verbs in the phrase

9.3 Log Likelihood Ratio

Similar to the conditional probability approach, the *log likelihood* approach also compares the number of times a certain kind of verb phrase has been judged as denoting work to the number of times it has occurred in the training corpus without being extracted. One advantage of the log likelihood ratio is

however that it also takes into account the number of times a specific token occurs in the training corpus, relative to other tokens, rendering a more fair score for low-frequency tokens as compared to high-frequency tokens. The log likelihood ratio (llr) is here calculated in accordance with the formula presented by Dunning [1993], defined as below:

	Event A	Everything but A
Event B	k ₁₁ : A + B	k ₁₂ : B only
Everything but B	k ₂₁ : A only	k ₂₂ : Neither A nor B

H = Shannon's entropy, computed as the sum of $(k_{ij} / \text{sum}(k)) \log(k_{ij} / \text{sum}(k))$
 $\text{llr} = 2 * \text{sum}(k) * (H(k) - H(\text{rowSums}(k)) - H(\text{colSums}(k)))$

Applied to the verb phrase ranking problem, the following values are used for the log likelihood variables in order to retrieve a ratio for the probability that a certain verb form denotes work:

- **k₁₁**
The number of times a specific verb occurs in the training corpus and is part of a phrase that the historians have defined as a phrase describing work.
- **k₁₂**
The number of times the same verb occurs in the training corpus without being extracted.
- **k₂₁**
The number of times any other verb occurs in the training corpus and is part of a phrase that the historians have defined as a phrase describing work.
- **k₂₂**
The number of times any other verb occurs in the training corpus without being extracted.

With the above formula, the log likelihood ratio is always given as a positive number. This means that a high number could either indicate a high probability that the phrase describes work, or a high probability that the phrase does **not** describe work. When applied to the ranking problem, I therefore also take into account the relative frequency with which the verb has been judged as describing work in the training corpus. If the verb in question occurs most frequently without being extracted, the log likelihood ratio is prefixed with a minus sign, and treated as representing the probability that the phrase at hand does not describe a working activity. In other cases, the probability score is

left as a positive number, thus representing the probability that the phrase at hand actually describes a working activity.

I try the following log likelihood settings applied to the ranking problem, where each setting is evaluated on purely tokenised (and normalised) word forms as well as on lemmatised data, yielding a total of twelve different settings:

words/lemmas The log likelihood ratio is calculated on the basis of all the tokens (or lemmas) in the phrase. The log likelihood score for the token/lemma with the highest log likelihood ratio is chosen as the ranking score for the whole phrase.

vb The log likelihood ratio is calculated solely on the basis of the verbs in the phrase. The likelihood score for the verb with the highest log likelihood ratio is chosen as the ranking score for the whole phrase.

vbcomp The log likelihood ratio is calculated separately for the verbs and for the non-verb tokens (or lemmas) in the verb phrase. The sum of the maximum verbal log likelihood and the maximum non-verbal log likelihood is chosen as the ranking score for the whole phrase. The hypothesis is that the verbal complements are of importance to distinguish in what contexts a certain verb describes a working activity. For intransitive verbs, only the maximum verbal log likelihood ratio is used for scoring.

vbcomp nn The log likelihood ratio is calculated as in the vbcomp setting, but for the non-verbal calculations, only the nouns are taken into account.

cooc The log likelihood ratio is calculated for the co-occurrence of the verb and each token (or lemma) in the complements. The maximum co-occurrence log likelihood is chosen as the ranking score for the whole phrase. For intransitive verbs, the maximum verbal log likelihood ratio is used for scoring.

cooc nn The log likelihood ratio is calculated as in the cooc setting, but only the nouns in the complements are accounted for.

9.4 Bag-of-Words Classification

In the *bag-of-words classification* approach, I run a support vector machine (SVM) classifier with the sequential minimal optimization (SMO) algorithm as defined by Platt [1998]. All experiments presented here are run with the default linear kernel SVM/SMO settings in the Weka data mining software package, version 3.6.10 [Hall et al., 2009]. I try three different feature selection models for the verb phrase ranking problem, where each model has been

applied both to normalised word forms and to lemmatised data, yielding a total of six different settings:

bag of words/lemmas Each word type (or lemma) occurring in the verb phrases in the training corpus is stored as a feature in the model. For every verb phrase to be ranked, each feature is then assigned a value of 1 or 0, depending on whether the specific word form (or lemma) represented by the feature is present in the phrase to be ranked or not.

bag of verbs In the bag-of-verbs setting, only those word forms (or lemmas) that the tagger has analysed as verbs are stored as features. Likewise, only word forms (or lemmas) in the phrase to be ranked that have been analysed as verbs will be compared towards the list of features.

bag of verbs and nouns The bag-of-verbs-and-nouns setting is similar to the bag-of-verbs setting, with the exception that both verbs and nouns are accounted for in this setting. The hypothesis is that the verbal complements, and in particular the nouns occurring in the complements, are of importance to distinguish in what contexts a certain verb describes a working activity.

10. Evaluation

Evaluation is performed by comparing the automatically extracted and ranked verb phrases to the manually classified verb phrases. It is however not a trivial task to decide which of the automatically extracted verb phrases that should be classified as denoting work, when comparing them to the gold standard of phrases extracted by the historians. Requiring the automatically extracted phrase to be identical to the manually extracted phrase would not be suitable, since the phrases extracted by the historians are not always phrases in the linguistic sense, but may include constituents that would normally be regarded as not belonging to the verb phrase, such as clause adverbials, non-adherent prepositional phrases, or the subject noun phrase. Likewise, the manually extracted segments sometimes exclude constituents that would normally be regarded as belonging to the verb phrase, such as indirect objects and adverbial complements. There are also inconsistencies in the spans of the manually extracted phrases, probably partly due to different excerptors, as exemplified in the previous chapter.

Similarly, the automatic extraction of verb phrases also results in incomplete verb phrases and phrases containing superfluous constituents. Still, since the overall aim of the verb phrase extraction process is to present elements in the text that may be of interest to the historians, partial phrases and phrases containing extra constituents would still point the user to the right text passage in the source material. Thus, both for training and evaluation I judge an automatically extracted verb phrase as describing work, if there is at least one verb in common between the automatically extracted phrase and the manual excerpt. This means that I run the risk of extracting the wrong instance and still judge it as correct, if there are several instances of the same verb form in the same case. This is especially true for frequent homonyms such as *ha* ('have'), which may be either a temporal auxiliary or a main verb and thus occur several times within the same case or even within the same sentence. In most cases, though, if the automatic excerpt has a verb in common with the manual excerpt, both phrases refer to the same instance. One authentic example from the Gender and Work database is the phrase *sålt een gårdh till hr Leijon Crona* ('sold a farm to Mr Leijon Crona'), which in the automatic excerpt is given as the shorter phrase *sålt een gårdh* ('sold a farm'), but will still be regarded as a true positive.

10.1 Evaluation Metrics

Three different evaluation metrics are applied to the verb phrase ranking results: *precision at k*, *R-precision*, and *average precision*, defined as below:

Precision at k is defined as the precision at certain positions in the list of ranked instances [Manning et al., 2008]. For example, precision at 10 is the precision achieved for the top-10 instances in the list. In my setup, I include precision at 10, 50, and 100 respectively.

R-precision (R-pre) is similar to precision at k, but requires a gold standard defining the total number of relevant instances. R-precision is then calculated by retrieving the precision score at the position in the list where the number of extracted verb phrases is equal to the number of relevant verb phrases in the gold standard. At this point, precision and recall are the same, which is why this measure is sometimes also referred to as the *break-even point* [Craswell, 2009]. R-precision can be summarised in the following formula:

R = number of relevant phrases in gold standard

r = extracted relevant phrases at position R

$$\mathbf{R}\text{-precision} = \frac{r}{R}$$

In my experiments, it is known that the total number of verb phrases denoting work in the evaluation part of the corpus is 1,254. Hence, R-precision is defined as precision at 1,254.

Average Precision (AVP) is calculated on the basis of the top n results in the extracted list, where n includes all positions in the list until all relevant instances have been retrieved [Zhang and Zhang, 2009]. The average precision can be expressed by the following formula:

r = rank for each relevant instance

P@r = precision at rank r

R = number of relevant phrases in gold standard

$$\mathbf{Average\ precision} = \frac{\sum_r P@r}{R}$$

10.2 Conditional Probability Results

Ranking based on conditional probability leads to a substantial improvement in the coverage of verbs denoting work among the top-listed instances, as compared to the baseline case, where the verb phrases are not ranked at all, but simply displayed in the order in which they appear in the source text. As shown in Table 10.1, not a single verb phrase describing work is among the

top-10 instances without ranking, and only 10% of the top-50 instances are phrases describing work. This could be compared to the token-based model using the maximum value for ranking, where eight out of the top-10 instances are true positives, and 66% of the top-50 instances denote work. At the break-even point (R-precision), nearly half of the positive instances are covered in this setting, as compared to only 23% without ranking. The average precision value follows the R-precision value closely for all settings.

	p@10	p@50	p@100	R-pre	AVP
Baseline	0.00	0.10	0.14	0.23	0.24
Token-based avg	0.50	0.66	0.63	0.46	0.44
Lemma-based avg	0.30	0.64	0.64	0.44	0.43
Token-based max	0.80	0.66	0.70	0.48	0.49
Lemma-based max	0.60	0.68	0.72	0.47	0.49

Table 10.1. Results for verb phrase ranking based on conditional probability. $p@10$ = precision at 10, $p@50$ = precision at 50, $p@100$ = precision at 100, $R\text{-pre}$ = R-precision, AVP = average precision, *Baseline* = results for the unranked list, *avg* = probability score based on average value, *max* = probability score based on maximum value.

The results also show that ranking based on the highest ranked verb for each phrase, rather than averaging over all the verbs, works the best. Furthermore, I had expected a positive effect of lemmatisation, but interestingly lemmatisation does not help much in the ranking process, and sometimes even lead to lower scores, especially for the models based on average. One reason could be that the kind of documents I am working with (court records and church documents) are almost exclusively written in the past tense, limiting the amount of different verb forms occurring for each lemma. There are also large groups of verbs denoting work, such as *köpa* ('to buy'), *sälja* ('to sell'), *arbeta* ('to work'), *tjäna* ('to serve) etc, that are so commonly occurring in the training and evaluation corpora that lemmatisation is of little help in the ranking process for these forms.

Despite the promising results, there is still room for improvement. The main problem with the conditional probability approach is that no consideration is taken to the number of times a specific verb occurs in the training corpus. Hence, if a certain verb occurs only once in the training corpus, and has been extracted by the historians, it will get the probability 1 of denoting work, and end up at the top of the list. This will be disadvantageous to verbs like *sell* or *buy* that occur many times in the corpus and are often, but not always, defined as describing work. Likewise, verbs occurring only once without being extracted will always end up at the bottom of the list, together with previously unseen verbs. This skewness is addressed by the log likelihood approach.

10.3 Log Likelihood Ratio Results

As presented in Table 10.2, the log likelihood approach, being more sophisticated in balancing the probabilities for low frequency word forms versus high frequency word forms, shows an improvement in the ranking results as compared to the conditional probability approach.

	p@10	p@50	p@100	R-pre	AVP
Baseline	0.00	0.10	0.14	0.23	0.24
Words	0.80	0.80	0.72	0.52	0.52
Lemmas	0.60	0.70	0.74	0.45	0.47
vb token-based	0.80	0.80	0.72	0.53	0.52
vb lemma-based	0.50	0.68	0.77	0.51	0.49
vbcomp token-based	0.80	0.84	0.83	0.46	0.49
vbcomp lemma-based	0.80	0.80	0.79	0.45	0.49
vbcomp nn token-based	0.90	0.82	0.78	0.53	0.52
vbcomp nn lemma-based	0.90	0.82	0.80	0.46	0.49
cooc token-based	0.90	0.76	0.81	0.36	0.42
cooc lemma-based	0.70	0.74	0.78	0.35	0.40
cooc nn token-based	0.50	0.76	0.77	0.31	0.35
cooc nn lemma-based	0.50	0.74	0.77	0.31	0.35

Table 10.2. Results for verb phrase ranking based on the log likelihood ratio. $p@10$ = precision at 10, $p@50$ = precision at 50, $p@100$ = precision at 100, $R\text{-pre}$ = R -precision, AVP = average precision, *Baseline* = results for the unranked list. See Section 9.3 for a description of the other abbreviations used in the table.

It is hard to tell which log likelihood setting is the best, since it depends on what evaluation metric is considered. One option would be to look closer at the results for precision at 100, since it would be a possible scenario to only display the top-100 instances to the user. From these results, it can be seen that the models where the complements are taken into account (*vbcomp* and *cooc* in the table) yield better results than the plain verb-based models. It is also clear that it is more successful to calculate the log likelihood for the verb and the complement separately, and return the sum of these values (*vbcomp*), than to compute a log likelihood score for the co-occurrence of the verb and any of the word forms in the complement (*cooc*).

Furthermore, higher precision at k results are achieved when the log likelihood score is calculated for all the word forms in the complement, than if only the nouns in the complement are considered (even though R -precision and average precision are slightly higher for the noun-restricted settings). A closer look at the top-ranked phrases reveal that they all include the indefinite article, e.g. *sålt en* ('sold a'), *köpt en* ('bought a'), *skjutit en* ('shot a'), *stulit en* ('stolen a'), etc. This is logical in a way, since it indicates that it is of

greater importance to the log likelihood ratio that **something** is sold or bought or worked with etc, than exactly **what** is sold or bought or worked with, where the latter would be better expressed by the nouns in the complement than by the indefinite article.

10.4 Bag-of-Words Classification Results

The ranking results for the bag-of-words classification approach are presented in Table 10.3.

	p@10	p@50	p@100	R-pre	AVP
Baseline	0.00	0.10	0.14	0.23	0.24
Words	0.60	0.88	0.84	0.49	0.53
Lemmas	0.50	0.82	0.81	0.49	0.52
vb token-based	1.00	0.92	0.87	0.52	0.55
vb lemma-based	1.00	0.94	0.85	0.50	0.53
vbnn token-based	0.80	0.92	0.91	0.50	0.54
vbnn lemma-based	0.70	0.92	0.88	0.50	0.54

Table 10.3. Results for verb phrase ranking based on machine learning. $p@10$ = precision at 10, $p@50$ = precision at 50, $p@100$ = precision at 100, $R\text{-pre}$ = R -precision, AVP = average precision, *Baseline* = results for the unranked list, *Words* = bag of words, *Lemmas* = bag of lemmas, *vb* = bag of verbs, *vbnn* = bag of verbs and nouns.

The results are generally higher than for both the conditional probability method and the log likelihood calculations. For the best precision at 100 results, 91% of the instances are verb phrases describing work. Similar to the results for conditional probability and log likelihood ratio, lemmatisation generally has no positive effect on the results. Unlike the results for the log likelihood approach though, it seems beneficial to exclude non-nouns from the complements in the machine learning approach. This is however only true for the precision at 100 metric, whereas the other metrics indicate the opposite.

10.5 Summary of the Results

Table 10.4 summarises the results for the methods with the highest precision at 100 score within the three different approaches. As seen from the table, the bag-of-words classification approach yields the highest score for every evaluation metric used, when comparing these results.

	p@10	p@50	p@100	R-pre	AVP
Baseline	0.00	0.10	0.14	0.23	0.24
Conditional Probability	0.60	0.68	0.72	0.47	0.49
Log Likelihood Ratio	0.80	0.84	0.83	0.46	0.49
Bag-of-words	0.80	0.92	0.91	0.50	0.54

Table 10.4. Summary of the results for verb phrase ranking. $p@10$ = precision at 10, $p@50$ = precision at 50, $p@100$ = precision at 100, $R\text{-pre}$ = R-precision, AVP = average precision, baseline = results for the unranked list.

11. Part III: Summary and Conclusion

In the third part of this thesis, I addressed the issue of information extraction from historical text. Due to the information need arisen within the Gender and Work project, the experiments were specifically focused on extraction of verb phrases describing work from Early Modern Swedish text. For this purpose I make use of input data that have been automatically normalised to a more modern spelling using techniques developed and evaluated in the first part of the thesis. Furthermore, the data is linguistically annotated using NLP tools developed for the modern language, as described in the second part of the thesis.

For the verb phrase extraction task, I tried three ranking approaches, aiming at presenting the verb phrases that are most likely to denote work at the top of the results list. The ranking approaches are based on 1) conditional probability, 2) log likelihood ratio, and 3) bag-of-words classification. Neither of the methods are dependent on semantically annotated data, since they all rely on binary classified training data of verb phrases describing work as opposed to verb phrases not describing work.

Although the ranking systems were trained on binary data rather than ranked data, all three methods yield very promising results. The bag-of-words classification approach reaches the highest scores according to all three evaluation metrics used (precision at k, R-precision, and average precision). The best bag-of-words setting is token-based (as opposed to lemma-based), taking both the verbs and the nouns in the verb phrases into account in the ranking process. In this setting, 91% of the top-100 instances in the results list are true positives.

Even though the experiments were conducted for the specific task of extracting and ranking verb phrases describing work in historical Swedish text, the methods developed are language-independent and could easily be applied to other languages and information needs by simply altering the training data. I believe this generic property of the information extraction pipeline to be an important aspect in order for this workflow to be of use for historians and other researchers in humanities. Hopefully, using this tool will enable researchers to search through larger volumes of historical text in less time, leaving more room for the actual analysis of the extracted data.

12. Conclusion

In recent years, large volumes of historical text have been digitised, enabling historians and other researchers in humanities to study various historical sources in a more systematic way, and to more easily search for text passages of interest to their research. However, there is still a lack of NLP tools adapted to historical text, which would facilitate the process of information extraction from these texts. The tools that do exist are generally developed for a specific language, time period and/or genre, and are often integrated in a larger system, making it hard to apply the tool to languages and information needs other than the intended ones. Even though there is a growing amount of historical texts available in an electronic format, researchers working with these texts are thus often still obliged to manually search through the text, in order to find the information they are interested in.

In this thesis I have presented a generic pipeline for information extraction from historical text. The aim of my work has been to develop a pipeline that should be applicable to historical text regardless of source language, genre, specific time period or information need. A further constraint is that the presented techniques should be modular and easily applicable, without being dependent on a full-fledged information extraction system. Furthermore, the information extraction pipeline should be applicable also in cases where little (or no) annotated historical data is available, as long as there is a set of basic NLP tools available for the modern language variant. This is an important aspect of my work, since linguistically annotated historical texts are rare, and lacking for many languages. This is also related to the fourth, and last, constraint stating that the linguistic analysis step should reuse existing state-of-the-art NLP tools, without adaptation or re-training.

My approach to information extraction from historical text consists of three modular steps: spelling normalisation, linguistic analysis, and information extraction. In the following, I discuss the results that I have achieved for each step of the pipeline, in relation to the research questions formulated in Section 1.1. The spelling normalisation step has been evaluated for several languages, time periods, and genres. The techniques for linguistic analysis and information extraction from historical text have also been designed to be generic and language-independent, but have mainly been evaluated in the context of the Gender and Work project. In this project, historians are interested in extracting information on what men and women did for a living in the Early Modern Swedish society (~1550–1800). This means that my experiments for these steps have been specifically focused on the extraction of verb phrases describing work from Early Modern Swedish text.

12.1 Spelling Normalisation

In the spelling normalisation step, the historical input text is automatically normalised to a more modern spelling, with the aim of making it possible to use contemporary taggers and parsers for the succeeding linguistic analysis step. In my thesis, I have developed and evaluated four different approaches to spelling normalisation:

1. rule-based normalisation
2. Levenshtein-based normalisation
3. memory-based normalisation
4. SMT-based normalisation

In the rule-based approach, normalisation rules are manually defined, based on known spelling changes from a language evolution perspective and/or based on spelling differences observed in text. This means that the rule-based normalisation approach may be applied also in cases where no annotated historical data is available. The main drawback of the rule-based approach is that it is highly language-dependent, and also time-consuming and knowledge-intense to implement. However, my experiments show that normalisation rules defined solely on the basis of a small sample of 17th century court records text, have a positive effect on normalisation accuracy for both older and younger texts, and for texts within the completely different genre of church documents.

In the Levenshtein-based normalisation approach, normalisation is treated as a spelling correction problem, which could be solved by edit distance comparisons between the historical spelling and word forms present in a modern dictionary. Thus, the only resource needed for implementing this normalisation approach is a modern language dictionary (or corpus) used as a basis for these edit distance calculations. This makes the Levenshtein-based method particularly suitable for languages with limited access to historical training data. It is also fairly trivial to apply the Levenshtein-based method to a new language, simply by substituting the modern language dictionary resource. If parallel texts in the form of historical word forms mapped to their modern spelling are available, these may be included in the Levenshtein-based approach to further improve normalisation accuracy, mainly by introducing weighted edit distance comparisons for frequently occurring spelling differences observed in the training data. The main drawback of the Levenshtein-based method is that it is very dependent on the coverage of the dictionary, since word forms that do not occur in the dictionary will be completely out of reach to this method. There are also cases where the word form with the smallest edit distance is not the intended one. One example from the English training corpus is the historical word form *stonde*, which is consequently normalised into *stone* by the Levenshtein-based method, instead of the intended *stand*, which has a larger edit distance to the original word form. If this hap-

pens for frequently occurring word forms in a language, this could have a huge impact on normalisation accuracy.

In the memory-based approach to spelling normalisation, a parallel corpus containing word forms in their historical and modern spellings is used as a basis for building a memory in which historical word forms are mapped to their previously normalised modern spellings. Each word form in the text to be normalised is matched against this memory, and whenever a word form is found in the memory, it is substituted by the most frequent normalisation candidate given in the dictionary. Unmatched word forms are left unchanged. This rather simplistic method turned out to work surprisingly well, but is obviously dependent on the size and content of the parallel corpus used as a memory.

In some respects, the Levenshtein-based method and the memory-based method have pros and cons that are complementary to each other. Whereas the Levenshtein-based method runs the risk of consequently normalising high-frequency word forms in an incorrect way, the memory-based method on the other hand has problems with less frequent word forms or spellings that are not covered in the parallel corpus used as a memory. To come to terms with the drawbacks of both approaches, I therefore tried a combination of the Levenshtein-based approach and the memory-based approach to spelling normalisation, in which only word forms that are not found in the memory are normalised using Levenshtein comparisons. This way, frequently occurring word forms are likely to be handled by the memory-based approach and run a lower risk of being normalised to an unintended word form, whereas previously unseen historical word forms are handled through Levenshtein comparisons. As expected, this hybrid approach to spelling normalisation improved normalisation accuracy, as compared to using any of the methods alone.

In the SMT-based approach, normalisation is treated as a translation task, where the historical word forms are translated into a modern spelling using statistical machine translation techniques. To handle differences in spelling rather than the full translation of words and phrases, translation is performed at a character-level, using the same kind of parallel training data as is used in the memory-based approach, that is historical word forms mapped to a modern spelling. A crucial advantage of the character-based SMT method as compared to traditional SMT models based on words and phrases, is that only small amounts of parallel training data are needed to achieve high-quality translations. This is due to the much smaller set of possible characters and character sequences to translate between, as compared to the possible combinations of words and phrases in a language. In my experiments, I showed that with a training corpus consisting of only 1,000 token pairs, a normalisation accuracy of 76.5% could be achieved, as compared to 83.9% when including the entire corpus of nearly 34,000 token pairs.

To assess the applicability of each normalisation approach to different source languages, I evaluated all methods except the rule-based method for five dif-

ferent languages: English, German, Hungarian, Icelandic, and Swedish. The results are not entirely conclusive. The SMT-based method with GIZA++ unigram alignment achieved the highest normalisation accuracy scores for English, German, and Swedish. Also for Hungarian, the SMT-based approach resulted in the best results, but with a bigram-based alignment model. For Icelandic on the other hand, the Levenshtein-based method combined with a normalisation memory outperformed all the other approaches. Since all approaches have a significant positive impact on normalisation accuracy, a main criterion for choosing an appropriate normalisation method for a new language would be what resources that are available for the particular language and time period. If parallel historical-modern training data is available, it could be worth the effort to train an SMT-based system for spelling normalisation, since this approach was the most successful one for four out of five languages.

12.2 Linguistic Analysis

In the linguistic analysis step, taggers and parsers available for the modern language are applied to the normalised text, enriching the input text with linguistic analysis labels to be used in the succeeding information extraction step. The method of using contemporary taggers and parsers for this task is in line with the aim of making the pipeline applicable to any language for which there are modern NLP tools available. It also complies with the criterion of not requiring access to annotated historical training data, which would be needed for training taggers and parsers specifically adapted to handling historical text.

In my work, the linguistic analysis step has been evaluated in the context of the specific information need arisen within the *Gender and Work* project at Uppsala University, where historians are storing information on what men and women did for a living in the Early Modern Swedish society (approximately 1550-1800). During this work, they have concluded that working activities often are expressed in the form of verb phrases, such as *to fish herring* or *to sell clothes*. Consequently, I evaluate tagging performance based on verb identification in Early Modern Swedish text. Likewise, parsing is evaluated based on the results for extracting complements to the verbs identified by the tagger, to build up a complete verb phrase. The results show a substantial improvement in both verb identification and verb phrase extraction when tagging and parsing is preceded by spelling normalisation. In fact, the results are rather close to the upper bound results achieved when applying the NLP tools to the text in its manually normalised gold spelling. This suggests that the spelling normalisation step is successful in enabling the use of modern NLP tools for linguistic analysis of historical text, and that issues other than spelling need to be taken into account in order to achieve even better results in the linguistic analysis phase.

One step in this direction is to make use of verb valency information in the verb phrase extraction process. I tried this in a post-processing step to the original verb phrase extraction module, deleting improbable complements suggested by the parser, and inserting complements suggested by the valency frame but not extracted by the parser. For this purpose, I make use of verb valency information present in contemporary valency dictionaries as well as in historical corpora. Evaluation results show a small but consistent improvement in both precision and recall when adding valency information to the extraction process. It is also interesting to note that adding valency frames extracted from the historical corpus shows larger improvement than adding valency frames extracted from contemporary valency dictionaries, even though the contemporary dictionaries cover more verb forms. One possible conclusion to draw from this could be that language changes over time, including the meaning of verbs, implying that contemporary verb valency frames do not necessarily hold for verbs occurring in older versions of the language.

12.3 Information Extraction

In the third and final step of my pipeline, the linguistically annotated text is used as a basis for information extraction, where textual units that are judged to be of interest to the task at hand are extracted from the text and ranked, so that those elements that are most likely to describe the specific information need in focus are displayed at the top of the results list.

As for the linguistic analysis step, the information extraction evaluation is also geared towards the specific information need expressed within the Gender and Work project, that is the task of extracting verb phrases describing work from Early Modern Swedish text. In line with the aim of making the pipeline generic, the methods developed for the information extraction step are however defined to be applicable to any language and information need, simply by replacing the training data. Furthermore, no semantically annotated data is needed for training, and I do not make use of a full-fledged information extraction system which would be hard to adapt to a different language or information need.

As training data in my experimental setup, I use a subset of the Gender and Work corpus, comprising phrases judged by the historians as describing work, with reference to the source text snippet that the phrase was extracted from. By doing spelling normalisation succeeded by tagging and parsing, as described in the previous sections, all phrases analysed as verb phrases in the source text are identified. From this analysis, both positive and negative instances are added to the training data. Positive instances are phrases analysed as verb phrases by the NLP tools and judged by the historians as describing work, whereas negative instances are all other phrases analysed as verb phrases by the NLP tools. Using these two data sets as training data, I tried three different

approaches to extraction and ranking of verb phrases describing work, based on:

1. conditional probability
2. log likelihood ratio
3. bag-of-words classification

In order not to favour methods that are biased towards a specific evaluation metric, I included three different metrics in the evaluation: precision at k, R-precision, and average precision. The results show that the bag-of-words classification outperforms the other methods for all three evaluation metrics. Furthermore, the results are very promising, showing that 91 out of the 100 top-ranked instances are true positives in the best-performing setting.

12.4 Summary of the Results

Figure 12.1 summarises my results for each step in the information extraction pipeline, when applied to the task of extracting verb phrases describing work from Early Modern Swedish text.

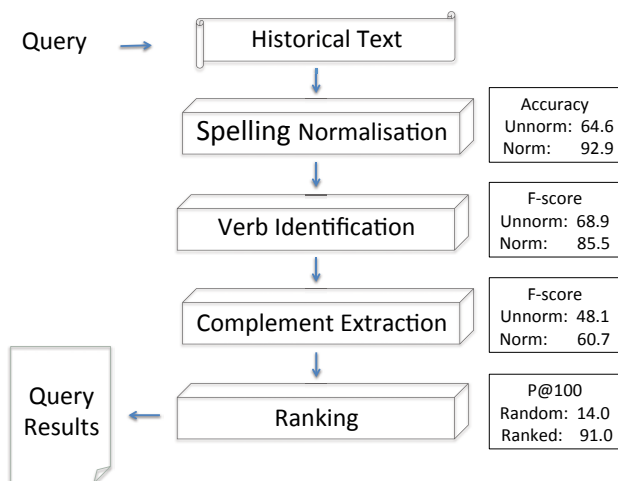


Figure 12.1. Summary of the results for each step in the information extraction pipeline, when applied to the task of extracting verb phrases describing work from Early Modern Swedish text. Unnorm = unnormalised input data, Norm = automatically normalised input data, Random = arbitrary ranked list of verb phrases, Ranked = list of verb phrases ranked using bag-of-words-classification, P@100 = precision at 100.

For the spelling normalisation step, normalisation accuracy scores are presented. In the original, unnormalised version of the text, approximately 64.6%

of the word forms in the evaluation corpus have a spelling that is identical to the spelling in the manually modernised gold standard version of the text. In the best-performing normalisation setting, that is the SMT-based method using GIZA++ with unigram models for character alignment, the proportion of word forms with a modern spelling increases to 92.9%. For the succeeding verb identification and complement extraction steps, a substantial improvement in F-score is achieved when tagging and parsing is preceded by spelling normalisation as compared to applying the NLP tools directly to the unnormalised text. Finally, the extracted verb phrases are ranked, with the aim of displaying those verb phrases that are likely to describe working activities at the top of the results list. In the figure, results are given in terms of precision at 100 for the best-performing setting, that is bag-of-words classification using unlemmatised verbs and nouns as features in the classification process. With arbitrary ranking, simply displaying the verb phrases in the order in which they occur in the text, only 14 out of the 100 verb phrases at the top of the results list are phrases classified as describing work. This could be compared to 91 out of 100 in the ranked list.

12.5 Future Work

Using the workflow presented in this thesis for information extraction from historical text shows very promising results in terms of traditional evaluation measures such as accuracy, precision, and recall. The most important aspect to evaluate is however how useful my pipeline is as an aid for historians and other researchers in humanities, in making the process of finding relevant information in historical text more efficient and the resulting excerpts more consistent. The historians in the Gender and Work project at the Department of History at Uppsala University have already tried out an early prototype of my pipeline, only including spelling normalisation succeeded by verb identification, with the possibility for the user to click on a verb to see all occurrences of this specific verb form in a larger context. The user feedback from this informal evaluation was very positive and encouraging for the further development of the verb phrase extraction and ranking modules. Currently, we are together with the historians planning a user evaluation of the complete information extraction pipeline, taking both quantitative and qualitative aspects into account. Quantitative aspects concern the number of relative instances that could be extracted from historical text within a certain period of time, when using the information extraction pipeline as an aid in the extraction process, as compared to performing a fully manual extraction. The qualitative aspects on the other hand, concern the actual information pieces extracted from the text. Are there differences in the phrases that are detected by the aid of automatic information extraction techniques, as compared to doing a fully manual extraction? One hypothesis is that the automatic system might favour ‘more of the same’, in the

sense that phrases that are similar to phrases already occurring in the training data will end up higher in the results list than phrases that also contain relevant information, but expressed in words or syntactic structures not occurring in the training data. On the other hand, the automatically extracted phrases could be expected to be more consistent, since different human excerptors tend to extract varying spans of a sentence for describing the same content, especially when there are no clear guidelines concerning what to consider as an appropriate information unit. Another quality aspect is that the tool will enable the user to examine all cases of a specific verb phrase construction at the same time, for example all instances of the phrase *hugga ved* ('chop wood'). This makes it likely that the user will be more consistent in the way these instances are classified, as compared to the manual setting in which different instances of the same construction are typically examined at different occasions.

From a more technical point of view, it would be interesting to try alternative methods for spelling normalisation, such as a phonetically based similarity measure for generating normalisation candidates. The lack of spelling conventions in the past often results in texts that are more similar to spoken language than to contemporary standard written language, meaning that phonetic similarity might in some cases be more adequate than orthographic similarity as a measure for comparison. There are also similarities between working with very old text and working with very modern text, such as chats and tweets. In both genres, there is a high degree of spelling variation, ad hoc abbreviations and ungrammatical structures imposing the problem of data sparseness. It would therefore be interesting to explore in more depth the methods used for linguistic analysis and information extraction from this kind of text.

Another idea is to try normalisation at a syntactic level, in addition to the current spelling normalisation which is performed at a word level only, without taking into account properties such as word order differences. In the field of machine translation, reordering techniques are commonly used for moving syntactic units to other positions within a sentence. Similar methods could potentially be used in the context of normalisation of historical text as well, where word order is often less strict than in modern language, or different from the modern language word order.

A third technical aspect to consider in order to obtain better linguistic analysis results would be to include language identification for automatic detection of text passages in which code switching occurs. This way, different sets of NLP tools could be applied to different parts of the text, depending on the source language suggested by the language identification tool.

Furthermore, for the linguistic analysis step, I have experimented on adding verb valency information in a post-processing step. An alternative way of making use of valency information would be to provide this information already in the parser training phase, by enriching the part-of-speech tags with information on whether a certain verb is likely to occur with for instance a particle or a prepositional complement. The hypothesis is that a parser trained

on this kind of data will be keen to search harder for the expected complements. It would also be interesting to explore the use of lexical semantics for identifying specific types of complements.

Future work also includes the application of my pipeline to more languages, time periods, genres, and information needs. As described above, I conducted and presented these kinds of experiments for the normalisation step, showing the generic properties of the proposed normalisation approaches. In this thesis, the linguistic analysis step and the information extraction step have however only been evaluated for the specific task of extracting verb phrases describing work from Early Modern Swedish text. I believe my methods to be generally applicable also to other information needs and textual properties, but I would like to confirm this through additional experiments. It could be mentioned that I already performed one such additional experiment. In Pettersson et al. [2013b], I evaluate the performance of the IceNLP tagger trained for contemporary Icelandic, when applied to 15th century Icelandic text. These experiments show an increase in tagging accuracy by 10 percentage points when normalisation is applied prior to tagging.

In a broader perspective, as the field of NLP for historical text is growing, the need for standardisation of historical language resources is increasing. In recent years, a number of historical corpora and databases have been created, such as the Penn Parsed Corpora of Historical English, the Icelandic Parsed Historical Corpus, the Tycho Brahe Parsed Corpus of Historical Portuguese, the Gold Standard Corpus of Early Modern German, and others. At present, there is however no generally accepted standard for how to annotate these kinds of resources. One step towards the standardisation of historical corpora is the Special Issue of the Language Resources and Evaluation Journal entitled *Converging Corpora: How to standardize historical corpora of typologically and genetically different languages*, to be published in 2016. The main questions addressed in this special issue are:

1. To what extent should the existing annotation schemes be extended for the incorporation of highly inflected languages?
2. How can existing schemes be extended to accomplish this?
3. How can the linguistic annotation of historical corpora be standardized to serve an easy-to-use data access for linguists?

A somewhat related issue concerns the need for comparability of results. In the NLP field in general, there are typically standard corpora and tools available (or under development) for many languages, such as the Wall Street Journal Corpus for English and the Stockholm-Umeå Corpus for Swedish. By using the same evaluation corpus for various experiments, the results achieved by different researchers are easily comparable to each other. Most work in the subfield of NLP for historical text is however conducted and evaluated on

data that is not comparable to the data used in other experiments by other researchers. Thus, the development of standard corpora and tools to be used in this specific subfield of NLP is called for.

Finally, both the establishment of annotation standards for historical corpora, and the creation of common corpora and tools to be used in the context of NLP for historical text, could gain from a cooperation between computational linguists and researchers in humanities (e.g. historical linguists). In general, I think that one of the most important future challenges is to bring the field of digital humanities closer to the field of NLP. As discussed in Chapter 2, a number of digital humanities centres are currently being established world-wide, to encourage this kind of cooperation. It is my belief that further cooperation between the disciplines would lead to fruitful interdisciplinary projects from which both fields would benefit.

References

- M. Ågren, R. Fiebranz, E. Lindberg, and J. Lindström. Making Verbs Count. The research project ‘Gender and Work’ and its methodology. *Scandinavian Economic History Review*, 59(3):271–291, 2011.
- A. Baron and P. Rayson. VARD2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, 2008.
- R. Basili and F. M. Zanzotto. Parsing Engineering and Empirical Robustness. *Natural Language Engineering*, 8:97–120, 2002.
- G. Bergman. *Kortfattad svensk språkhistoria*. Prisma Magnum, Stockholm, 5th edition, 1995.
- A. Bergs, editor. *English Historical Linguistics*, volume 1. Walter de Gruyter, 2012.
- D. Biber, E. Finegan, and D. Atkinson. ARCHER and its challenges: Compiling and Exploring A Representative Corpus of Historical English Registers. In *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, 1993*, pages 1–13. 1994.
- A. Bilal. Lexical Normalisation of Twitter Data. *Computing Research Depository, CoRR*, 2014. URL <http://arxiv.org/abs/1409.4614>.
- K. Bjarnadóttir. The Database of Modern Icelandic Inflection. In *AfLaT2012/SALTMIL Joint Workshop on Language Technology for Normalisation of Less-Resourced Languages*, pages 13–18, 2012.
- A. W. Black and P. Taylor. The Festival Speech Synthesis System: system documentation. Technical report, Human Communication Research Centre, University of Edinburgh, Scotland, UK, 1997. URL <http://www.cstr.ed.ac.uk/projects/festival/>.
- M. Bollmann. (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, pages 3–14, 2012.
- M. Bollmann. POS Tagging for Historical Texts with Sparse Training Data. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 11–18, 2013.
- M. Bollmann, F. Petran, and S. Dipper. Rule-Based Normalization of Historical Texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage (LaTeCH)*, pages 34–42, 2011.
- L. Borin, D. Kokkinakis, and L.-J. Olsson. Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, pages 1–8, 2007.
- L. Borin, M. Forsberg, and L. Lönnegren. SALDO 1.0 (Svenskt associationslexikon version 2). Språkbanken, University of Gothenburg, 2008.

- G. Bouma and Y. Adesam. Experiments on sentence segmentation in Old Swedish editions. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*, NEALT Proceedings Series 18, pages 11–26, 2013.
- T. Brants. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, pages 224–231, 2000.
- F. Braune and A. Fraser. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of COLING, Poster Volume*, pages 81–89, 2010.
- P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- L. Campbell. *Historical Linguistics*. Edinburgh University Press, 2013.
- N. Craswell. R-Precision. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 2453–2453. Springer US, 2009. URL http://dx.doi.org/10.1007/978-0-387-39940-9_486.
- C. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. The Szeged Treebank. In *Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD)*, pages 123–131, 2005.
- T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- N. Edling. *Uppländska domböcker*. Almqvist & Wiksells, 1937.
- E. Ejerhed and G. Källgren. Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University, 1997.
- M. Federico, N. Bertoldi, and M. Cettolo. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech 2008*, pages 1618–1621, 2008.
- C. Galves and P. Faria. Tycho Brahe Parsed Corpus of historical Portuguese, 2010. URL <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>.
- P. Gardner-Chloros. *Code-switching*. Cambridge University Press, 2009.
- P. Halácsy, A. Kornai, and C. Oravecz. HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 209–212, 2007.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11:1:10–18, 2009.
- B. Han and T. Baldwin. Lexical Normalisation of Short Text Messages: Mkn Sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 368–378, 2011.
- S. Helgadóttir, A. Svavarsdóttir, E. Rögnvaldsson, K. Bjarnadóttir, and H. Loftsson. The Tagged Icelandic Corpus (MÍM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, pages 67–72, 2012.
- S. Jiampojamarn, G. Kondrak, and T. Sherif. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. In *Proceedings of the Annual Conference of the North American Chapter of the Association for*

- Computational Linguistics (NAACL-HLT 2007)*, pages 372–379, 2007.
- B. Jurish. Finding canonical forms for historical German text. In A. Storrer, A. Geyken, A. Siebert, and K.-M. Würzner, editors, *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing (KONVENS 2008)*, pages 27–37. Mouton de Gruyter, Berlin, 2008.
- P. Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, 2007.
- G. Kondrak and B. Dorr. Identification of confusable drug names. In *Proceedings of COLING 2004*, pages 952–958, 2004.
- A. Kroch and A. Taylor. Penn-Helsinki Parsed Corpus of Middle English, 2000.
- A. Kroch, B. Santorini, and A. Dierani. Penn-Helsinki Parsed Corpus of Early Modern English, 2004.
- K. Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439, 1992.
- C. Lehmann. Directions for interlinear morphemic translations. *Folia Linguistica*, 16:199–224, 1982.
- W. P. Lehmann. *Historical Linguistics*. Routledge, 1992.
- V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.
- A. Loth, editor. *Late Medieval Icelandic Romances I*. Kaupmannahöfn, Copenhagen, 1962.
- C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- M. Markus. *Manual of ICAMET (Innsbruck Computer Archive of Machine-Readable English Texts)*. Leopold-Franzens-Universität Innsbruck, 1999.
- D. Matthews. Machine Transliteration of Proper Names. Master’s thesis, School of Informatics, 2007.
- B. B. Megyesi. The Open Source Tagger HunPoS for Swedish. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA)*, pages 239–241, 2009.
- P. Nakov and J. Tiedemann. Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, 2012.
- J. Nivre. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*, 34(4):513–553, 2008.
- J. Nivre, J. Hall, and J. Nilsson. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)*, pages 2216–2219, 2006a.
- J. Nivre, J. Nilsson, and J. Hall. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)*, pages 24–26, 2006b.

- F. J. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.
- L. Padró and E. Stanilovsky. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC) ELRA*, pages 2473–2479, 2012.
- H. Palsson, editor. *The Uppsala Edda*. Viking Society for Northern Research, 2012.
- M. Pennacchiotti and F. M. Zanzotto. Natural Language Processing Across Time: An Empirical Investigation on Italian. In *Advances in Natural Language Processing. GoTAL, LNAI*, volume 5221, pages 371–382, 2008.
- D. L. Pennell and Y. Liu. A Character-Level Machine Translation Approach for Normalization of SMS Abbreviations. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 8–13, 2011.
- E. Pettersson and J. Nivre. Automatic Verb Extraction from Historical Swedish Texts. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 87–95, 2011.
- E. Pettersson, B. Megyesi, and J. Nivre. Rule-Based Normalisation of Historical Text - a Diachronic Study. In *Proceedings of the First International Workshop on Language Technology for Historical Text(s)*, pages 333–341, 2012a.
- E. Pettersson, B. Megyesi, and J. Nivre. 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 (LaTeCH)*, pages 65–74, 2012b.
- E. Pettersson, B. Megyesi, and J. Nivre. Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 163–179. Linköping Electronic Conference Proceedings 85, 2013a.
- E. Pettersson, B. Megyesi, and J. Tiedemann. An SMT Approach to Automatic Annotation of Historical Text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA*, NEALT Proceedings Series 18, pages 54–69. Linköping Electronic Conference Proceedings 87, 2013b.
- E. Pettersson, B. Megyesi, and J. Nivre. A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 32–41, Gothenburg, Sweden, 2014a.
- E. Pettersson, B. Megyesi, and J. Nivre. Verb Phrase Extraction in a Historical Context. In *The First Swedish National SWE-CLARIN Workshop at the Swedish Language Technology Conference (SLTC)*, Uppsala, Sweden, 2014b.
- E. Pettersson, B. Megyesi, and J. Nivre. Improving Verb Phrase Extraction from Historical Text by use of Verb Valency Frames. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 153–162, 2015a.
- E. Pettersson, B. Megyesi, and J. Nivre. Ranking Relevant Verb Phrases Extracted from Historical Text. In *Proceedings of the 9th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 39–47, 2015b.
- M. Piotrowski. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers, 2012.

- J. C. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical report, Advances in Kernel Methods - Support Vector Learning, 1998.
- P. Rayson. Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora. In *Proceedings of the Corpus Linguistics Conference: CL2007*, University of Birmingham, UK, 2007. UCREL. URL http://eprints.lancs.ac.uk/13011/1/192_Paper.pdf.
- P. Rayson, D. Archer, and S. Nicholas. VARD versus Word – A comparison of the UCREL variant detector and modern spell checkers on English Historical Corpora. In *Proceedings from the Corpus Linguistics Conference Series on-line e-journal*, volume 1, Birmingham, UK, July 2005. URL http://eprints.lancs.ac.uk/12686/1/cl2005_varword.pdf.
- D. Ringe and J. F. Eska. *Historical Linguistics - Toward a Twenty-First Century Reintegration*. Cambridge University Press, 2013.
- V. Rocio and P. J. G. Lopes. An infra-structure for diagnosing causes for partially parsed natural language input. In *Proceedings of the Sixth International Symposium on Social Communication*, pages 550–554, 1999.
- V. Rocio, M. A. Alves, J. G. Lopes, M. F. Xavier, and G. Vicente. Automated Creation of a Partially Syntactically Annotated Corpus of Medieval Portuguese Using Contemporary Portuguese Resources. In *Proceedings of the ATALA workshop on Treebanks*, pages 59–67, 1999.
- E. Rögnvaldsson, A. K. Ingason, E. F. S. sson, and J. Wallenberg. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 1977–1984, 2012.
- E. Rögnvaldsson and S. Helgadóttir. Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In C. Sporleder, A. van den Bosch, and K. Zervanou, editors, *Language Technology for Cultural Heritage, Theory and Applications of Natural Language Processing*, pages 63–76. Springer Berlin Heidelberg, 2011.
- J. Salmons. *A History of German – What the past reveals about today’s language*. Oxford University Press, 2012.
- C. Sánchez-Marco, G. Boleda, and L. Padró. 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 (LaTeCH)*, pages 1–9, 2011.
- S. Scheible, R. J. Whitt, M. Durrell, and P. Bennett. A Gold Standard Corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 124–128, 2011a.
- S. Scheible, R. J. Whitt, M. Durrell, and P. Bennett. 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)*, pages 19–23, 2011b.
- H. Schendl. Linguistic aspects of code-switching in Medieval English texts. In D. Trotter, editor, *Multilingualism in Later Medieval Britain*, pages 77–92. Boydell & Brewer, 2002.

- H. Schendl and L. Wright. *Code-switching in Early English*. De Gruyter Mouton, 2011.
- Y. Scherrer and T. Erjavec. Modernizing historical Slovene words with character-based SMT. In *Proceedings of the 4th Biennial Workshop on Balto-Slavic Natural Language Processing*, pages 58–62, 2013.
- H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, 1994.
- J. Schmied. The Lampeter Corpus of Early Modern English Tracts. In *Corpora Across the Centuries*, pages 199–211. Amsterdam, 1994.
- G. Schneider. *Hybrid long-distance functional dependency parsing*. PhD thesis, University of Zürich, Faculty of Arts, 2008.
- G. Schneider. Adapting a parser to historical English. In J. Tyrkkö, M. Kilpiö, T. Nevalainen, and M. Rissanen, editors, *Studies in Variation, Contacts and Change in English*, volume 10. Helsinki: VARIENG, 2012. URL <http://www.helsinki.fi/varieng/series/volumes/10/schneider/>.
- E. Simon. Corpus building from Old Hungarian codices. In *The Evolution of Functional Left Peripheries in Hungarian Syntax*, pages 224–236. Oxford University Press, 2014.
- S. Stymne. German Compounds in Factored Statistical Machine Translation. In *Proceedings of GoTAL, 6th International Conference on Natural Language Processing*, pages 464–475, 2008.
- S. Stymne and M. Holmqvist. Processing of Swedish Compounds for Phrase-Based Statistical Machine Translation. In *Proceedings of the 12th EAMT conference*, pages 182–191, 2008.
- W. Teubert. German Parole Corpus. Electronic resource, Oxford Text Archive, 2003.
- J. Tiedemann. *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. PhD thesis, Uppsala University, Uppsala, Sweden, 2003. Anna Sångvall Hein, Åke Viberg (eds): *Studia Linguistica Upsaliensia*.
- J. Tiedemann. Character-based PSMT for closely related languages. In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 12–19, 2009.
- J. Tiedemann and P. Nabende. Translating Transliterations. *International Journal of Computing and ICT Research*, 3(1):33–41, 2009.
- E. van Gelderen. *History of the English Language*. John Benjamins Publishing Company, 2006.
- D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Proceedings of the RANLP*, pages 590–596, 2005.
- D. Vilar, J.-T. Peter, and N. Hermann. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, 2007.
- E. Zhang and Y. Zhang. Average Precision. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 192–193. Springer US, 2009.

Appendix A.

Early Modern Swedish Normalisation Rules

Below, I list the set of 29 hand-written rules for normalisation of Early Modern Swedish text, described and evaluated in Section 3.2. The rules are based partly on the reformed Swedish spelling introduced in 1906 [Bergman, 1995], and partly on the initial 20 sentences (984 tokens) in *Per Larssons dombok*, a court records text from 1638 [Edling, 1937]. The rules are given in their raw Perl script form, with examples of word forms targeted by each rule.

```
## 1) qvarnar -> kvarnar (1906 spelling reform)
    $line=~s/qv/kv/g;
    $line=~s/Qv/Kv/g;

## 2) hvar, hvad, wore -> var, vad, vore (1906 spelling reform)
    $line=~s/^(hv|w)/v/g;
    $line=~s/^(Hv|W)/V/g;

## 3) hafuer, gifuess, lefuer, haffuer, hafva ->
##     haver, givess, lever, haver, hava
    $line=~s/($vowel)ff?[uv]($vowel)/$1v$2/g;

## 4) fördärfvat, blijfva -> fördärvat, bliva
    $line=~s/j?fv/$1v/g;

## 5) een, saaken, söokia -> en, saken, sökia
    $line=~s/($vowel)\1/$1/g;

## 6) ähr, ahntaga -> är, antaga
    $line=~s/($vowel)h([nr])/1$2/g;

## 7) uthskickadhe, uthfhöra, sägher ->
##     utskickade, utföra, säger
    $line=~s/([dtfgk])h/$1/g;

## 8) elliest, bevilliat -> eljest, beviljat
    $line=~s/lli([ae])/lj$1/g;
```

```

## 9) varidt -> varit (1906 spelling reform)
    $line=~s/födt/fött/g;
    $line=~s/dt/t/g;

## 10) dömbt, dömbdes, benämbdh -> dömt, dömdes, benämnd
    $line=~s/m[bp]t/mt/g;
    $line=~s/m[bp]d/mnd/g;

## 11) slogz, skötz -> slogs, sköts
    $line=~s/z/s/g;

## 12) försöria(s) -> försörja(s)
    $line=~s/([a])ria([t])/$1rja$2/g;

## 13) vijka, bevijsa, blijva -> vika, bevisa, bliva
    $line=~s/iji/j/g;
    $line=~s/ij/i/g;

## 14) häfdar -> hävdar
    $line=~s/($vowel)fd/$1vd/g;

## 15) föregaf -> föregav
    $line=~s/gaff?($|\s)/gav$1/;

## 16) blef -> blev
    $line=~s/eff?($|\s)/ev$1/;

## 17) affsagt -> avsagt
    $line=~s/^[Aa]ff/$1v/;

## 18) schall -> skall
    $line=~s/sch/sk/;

## 19) kiöpt -> köpt
    $line=~s/kiö/kö/;

## 20) prätenderat -> pretenderat
    $line=~s/(æ|ae)/e/g;

## 21) ehrläggia, sökia -> erlägga, söka
    $line=~s/(gg|k)ia/$1a/g;

## 22) huilken -> vilken
    $line=~s/hui/vi/g;

```

```

## 23) avsachnat -> avsaknat
## (restricted to succeeding consonants to avoid overgeneration
## for words such as 'och', 'doch' etc.)
    $line=~s/ch($cons)/k$1/g;

## 24) giort -> gjort
    $line=~s/~giort($|\s)/gjort$1/;

## 25) voro -> vore
    $line=~s/~voro($|\s)/vore$1/;

## 26) effter, ofta -> efter, ofta
    $line=~s/fft/ft/;

## 27) givess, avdömass, prövass -> gives, avdömas, prövas
    $line=~s/ss($|\s)/s$1/g;

## 28) af -> av
    $line=~s/([Aa])f([\^ft])/$1v$2/g;

## 29) iemte, sielf -> jämte, själf
    $line=~s/ie(l|r|mt)/jä$1/g;
    $line=~s/Ie(l|r|mt)/Jä$1/g;
    $line=~s/(T|t)ien/$1jän/g;

```


ACTA UNIVERSITATIS UPSALIENSIS
Studia Linguistica Upsaliensia
Editors: Joakim Nivre and Åke Viberg

1. *Jörg Tiedemann*, Recycling translations. Extraction of lexical data from parallel corpora and their application in natural language processing. 2003.
2. *Agnes Edling*, Abstraction and authority in textbooks. The textual paths towards specialized language. 2006.
3. *Åsa af Geijerstam*, Att skriva i naturorienterande ämnen i skolan. 2006.
4. *Gustav Öquist*, Evaluating Readability on Mobile Devices. 2006.
5. *Jenny Wiksten Folkeryd*, Writing with an Attitude. Appraisal and student texts in the school subject of Swedish. 2006.
6. *Ingrid Björk*, Relativizing linguistic relativity. Investigating underlying assumptions about language in the neo-Whorfian literature. 2008.
7. *Joakim Nivre, Mats Dahllöf and Beáta Megyesi*, Resourceful Language Technology. Festschrift in Honor of Anna Sågvall Hein. 2008.
8. *Anju Saxena & Åke Viberg*, Multilingualism. Proceedings of the 23rd Scandinavian Conference of Linguistics. 2009.
9. *Markus Saers*, Translation as Linear Transduction. Models and Algorithms for Efficient Learning in Statistical Machine Translation. 2011.
10. *Ulrika Serrander*, Bilingual lexical processing in single word production. Swedish learners of Spanish and the effects of L2 immersion. 2011.
11. *Mattias Nilsson*, Computational Models of Eye Movements in Reading : A Data-Driven Approach to the Eye-Mind Link. 2012.
12. *Luying Wang*, Second Language Acquisition of Mandarin Aspect Markers by Native Swedish Adults. 2012.
13. *Farideh Okati*, The Vowel Systems of Five Iranian Balochi Dialects. 2012.
14. *Oscar Täckström*, Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision. 2013.
15. *Christian Hardmeier*, Discourse in Statistical Machine Translation. 2014.
16. *Mojgan Seraji*, Morphosyntactic Corpora and Tools for Persian. 2015.
17. *Eva Pettersson*, Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. 2016.

