Parsing the Past – Identification of Verb Constructions in Historical Text

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

Abstract

Even though NLP tools are widely used for contemporary text today, there is a lack of tools that can handle historical documents. Such tools could greatly facilitate the work of researchers dealing with large volumes of historical texts. In this paper we propose a method for extracting verbs and their complements from historical Swedish text, using NLP tools and dictionaries developed for contemporary Swedish and a set of normalisation rules that are applied before tagging and parsing the text. When evaluated on a sample of texts from the period 1550–1880, this method identifies verbs with an F-score of 77.2% and finds a partially or completely correct set of complements for 55.6% of the verbs. Although these results are in general lower than for contemporary Swedish, they are strong enough to make the approach useful for information extraction in historical research. Moreover, the exact match rate for complete verb constructions is in fact higher for historical texts than for contemporary texts (38.7% vs. 30.8%).

1 Introduction

Today there is an abundance of NLP tools that can analyse contemporary language and extract information relevant to a particular user need, but there is a real lack of tools that can handle historical documents. Historians and other researchers working with older texts are still mostly forced to manually search large amounts of text in order to find the passages of interest to their research. Developing tools to facilitate this process is a great challenge, however, as historical texts vary greatly in both spelling and grammar between different authors, genres and time periods, and even within the same text, due to the lack of spelling conventions. In addition to this, there is a shortage of annotated resources that can be used for the development and evaluation of new tools.

The work presented in this paper has been carried out in cooperation with historians, who are studying what men and women did for a living in the Early Modern Swedish society. Currently, the historians are manually extracting segments describing work activities from a number of historical texts, and entering them into a database, the Gender and Work Database. Their work so far has shown that the typical segment describing an activity of work is a verb construction, that is, a verb together with its complements (Ågren et al., 2011). (Examples of such segments can be found below in Table 1.) It is very likely that the manual effort and the time needed by the historians to find these segments and enter them into the database could be substantially reduced if verbs and their complements were automatically extracted and presented to the historian. This would give a general overview of the content of a text, and the task of the historian would be to select those segments that are actually describing work activities. By linking extracted segments back to larger passages of text, historians would also be able to find additional segments that were missed by the first automatic extraction. The core of such a system would be a component for identifying verb constructions in running text.

We propose a method for automatically identifying verbs and their complements in various types of historical documents, produced in the Early Modern Swedish period (1550–1800). The method is based on using existing NLP tools for

contemporary Swedish, in particular a part-of-speech tagger and a syntactic parser, and automatically normalising the input text into a more modern orthography before running it through the tagger and parser. In order to increase the precision of complement extraction, we use valency dictionaries to filter out unlikely complements in the output of the syntactic parser. Using this method, we are able to identify verbs with an F-score of 77.2% and find a partially or completely correct set of complements for 55.6% of the verbs. To our knowledge, extracting verb constructions from historical texts is a task that has not been directly addressed in previous research, which means that these results are also important in setting benchmarks for future research.

The paper is structured as follows. Section 2 reviews related work. Section 3 describes the method for identification of verbs and complements in more detail. Section 4 presents the data and evaluation metrics used in our experiments, and Section 5 discusses the results of the evaluation. Finally, Section 6 concludes the paper.

2 Related Work

Using NLP tools for analysing historical texts is still to a large extent unexplored. There is however a growing interest in this area, and there have been attempts to analyse historical texts (1) by using contemporary NLP tools as they are, (2) by using such tools in combination with normalisation rules and/or dictionaries covering historical language variation, and (3) by training new tools on annotated historical corpora.

Pennacchiotti and Zanzotto (2008) concluded that contemporary NLP tools are not suitable as they are for analysing historical text. They tried to use a contemporary dictionary, morphological analyser and part-of-speech tagger to analyse Italian texts from the period 1200–1881. In their experiments, the dictionary only covered approximately 27% of the words in the oldest text, as compared to 62.5% of the words in a contemporary Italian newspaper text. Consequently, the morphological analyser based on the dictionary reached an accuracy of only 48%, as compared to 91% for contemporary text. Similarly, the part-of-speech tagger used reached an accuracy of only 54%, as compared to 97% for contemporary text.

Oravecz et al. (2010) included a standardisation/normalisation step in their work on semi-automatically annotating a corpus of Old Hungarian. Normalisation was performed using a noisy channel model combined with morphological analysis filtering and decision tree reranking. Combining these methods, they reached a normalisation precision of 73.3%.

Rocio et al. (1999) used a grammar of contemporary Portuguese to syntactically annotate medieval Portuguese texts. A dictionary and inflectional rules for medieval Portuguese were added to the parser, to make it suitable for handling these texts. This approach proved to be successful for partial parsing of medieval Portuguese texts, even though there were some problems remaining concerning grammar limitations, dictionary incompleteness and insufficient part-of-speech tagging.

Sánchez-Marco et al. (2011) adapted an existing NLP tool to deal with Old Spanish. The adapted tool had an accuracy of 94.5% in finding the right part of speech, and 89.9% accuracy in finding the complete morphological tag. The adaptation was performed on the basis of a 20 million token corpus of texts from the 12th to the 16th century, and included expansion of the dictionary, modification of tokenisation and affixation rules, and retraining of the tagger. The retraining was based on a gold standard of 30,000 tokens, where the tokens were first pre-annotated with the contemporary tagger, and then manually corrected. Adding new words to the dictionary had the highest impact on the results. This was done by automatically generating word forms through mapping old spelling variants to their contemporary counterparts.

Pettersson and Nivre (2011) presented a study on automatically extracting verbs from Swedish 17th century texts, using contemporary language technology tools combined with normalisation of the input text. The verb extraction process included an iterative process of normalisation and morphological analysis, followed by part-of-speech tagging for disambiguation of competing interpretations and for analysing words still unknown to the morphological analyser after all normalisation rules had been applied. Using this method, verbs were extracted with 82% precision and 88% recall. The study also included the results of using only the part-of-speech tagger for verb recognition, i.e., dropping the morphological analyser. This resulted in a small decrease in precision to 81% and in recall to 86%.

3 Extraction of Verb Constructions

In this paper, we will focus on adapting existing NLP tools by adding normalisation rules prior to processing. We will mainly follow the methodology for verb extraction described in Pettersson and Nivre (2011), but adding the extraction of not only the verbs themselves, but also their adherent complements. It would perhaps have been desirable to use tools specifically trained for analysing historical texts. This would however be a resource-demanding task, considering the lack of annotated data and the language variation, and is currently not a realistic scenario.

The goal is to automatically extract verbs and relevant complements from historical texts, in order to give an overview of the contents and present segments that are possibly describing work activities. In this context, we use the term *complement* in a broad sense and do not impose a sharp distinction between complements and adjuncts, especially not for prepositional phrases. This is motivated by the fact that in the Gender and Work Database, both segments that would traditionally be seen as complements and phrases that would rather be categorised as adjuncts have been considered relevant.

A closer look at the database shows that 67% of the entered segments consist of a verb with a direct object. Other common constructions are verbs with a prepositional complement (11%), verbs with both a direct object and a prepositional complement (10%), and (intransitive) verbs without complements (7%). Table 1 illustrates the most common construction types found in the database, which have been used to define the rules for extracting complements from parsed sentences. There were also a small number of segments (8 in total), that we were not able to categorise.

Freq	Comp	Source Text Example
273	dobj	*dhe bärgadhe [Höö]* they harvested [Hay]
47	pcomp	*[med een häst] kiörtt* driven [with a horse]
43	dobj + pcomp	*[det kiöpet] Han* *[med hänness man] giort* [the bargain] He made [with her husband]
30	intrans	*mala* to grind
5	dobj + infc	*hulpit [Muremest:]* *[inläggia Trappestenar]* helped [the Bricklayer] [to make a Stone Stair]
3	indobj + dobj	*[honom] [Järnet] sålltt* sold [him] [the Iron]
1	subc	*tillsee [att icke barnen skulle göra skada]* see to it [that the children do not do any harm]

Table 1: Segments describing work activities in the Gender and Work Database; verbs underlined; complements in brackets. Grammatical functions: dobj = direct object, pcomp = prepositional complement, intrans = intransitive, indobj = indirect object, infc = infinitive clause, subc = subordinate clause.

3.1 System Overview

The extraction of verbs and their complements is basically performed in five steps:

1. Tokenisation
2. Normalisation
3. Part-of-speech tagging
4. Parsing
5. Extraction of verb constructions

Tokenisation is performed with a simple tokeniser for Swedish that has not been adapted for historical texts.

3.2 Normalisation

After tokenisation, each word is normalised to a more modern spelling using a set of 29 handcrafted rules. The rules were developed using a text sample from *Per Larssons dombok*, a selection of court records from 1638 (Edling, 1937), a sample that has not been used in subsequent evaluation. An example of a normalisation rule is the transformation of the letter combination *sch* into *sk*, as in the old spelling *schall* that is normalised to the contemporary spelling *skall* ("shall/should"). Some additional rules were also formulated based on the reformed Swedish spelling introduced in 1906 (Bergman, 1995). This set of rules includes the transformation of double vowels into a single vowel, as in *sööka*, which is normalised into *söka* ("search").

3.3 Part-of-speech Tagging

The purpose of part-of-speech tagging in our experiments is both to find the verbs in the text and to prepare for the parsing step, in which the complements are identified. Part-of-speech tagging is performed using HunPOS (Halácsy et al., 2007), a free and open source reimplementation of the HMM-based TnT-tagger by Brants (2000). The tagger is used with a pre-trained language model based on the Stockholm-Umeå Corpus (SUC), a balanced, manually annotated corpus of different text types representative of the Swedish language in the 1990s, comprising approximately one million tokens (Gustafson-Capková and Hartmann, 2006). Megyesi (2009) showed that the HunPOS tagger trained on SUC, is one of the best performing taggers for (contemporary) Swedish texts.

3.4 Parsing

The normalised and tagged input text is parsed using MaltParser version 1.6, a data-driven dependency parser developed by Nivre et al. (2006a). In our experiments, the parser is run with a pre-trained model[1] for parsing contemporary Swedish text, based on the Talbanken section of the Swedish Treebank (Nivre et al., 2006b). The parser produces dependency trees labeled with grammatical functions, which can be used to identify different types of complements.

3.5 Extraction of Verb Constructions

The extraction of verb constructions from the tagged and parsed text is performed in two steps:

1. Every word form analysed as a verb by the tagger is treated as the head of a verb construction.

2. Every phrase analysed as a dependent of the verb by the parser is treated as a complement provided that it has a relevant grammatical function.

The following grammatical functions are defined to be relevant:

1. Subject (SS)

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)

Subjects are included only 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.

In an attempt to improve precision in the complement extraction phase, we also use valency dictionaries for filtering the suggested complements. The valency frame of a verb tells us what complements the verb is likely to occur with. The assumption is that this information could be useful for removing unlikely complements, i.e., complements that are not part of the valency frame for the verb in question. The following example illustrates the potential usefulness of this method:

J midler tijd kom ***greffuinnans gotze*** *fougte thijtt*
However, **the Countess' estate** bailiff came there

In this case, the parser analysed the partial noun phrase *greffuinnans gotze* ("the Countess' estate") as a direct object connected to *kom* ("came"). However, since the word *kom* is present in the valency dictionaries, we know that it is an intransitive verb that does not take a direct object. Hence, this complement can be removed. The valency dictionaries used for filtering are:

1. The Lexin dictionary, containing 3,550 verb lemmas with valency information.[2]

2. The Parole dictionary, containing 4,308 verb lemmas with valency information.[3]

3. An in-house machine translation dictionary, containing 2,852 verb lemmas with valency information.

4 Experimental Setup

4.1 Data

In order to evaluate the accuracy of our method, we have used ten texts from the period 1527–1737. The text types covered are court records

[1] http://maltparser.org/mco/swedish_parser/swemalt.html

[2] http://spraakbanken.gu.se/lexin/valens_lexikon.html

[3] http://spraakdata.gu.se/parole/lexikon/swedish.parole.lexikon.html

and documents related to the Church. In total, there are 444,075 tokens in the corpus, distributed as follows (number of tokens in parentheses):

Court records:

1. *Per Larssons dombok* (subset), 1638 (11,439)
2. *Hammerdals tingslag*, 1649–1686 (66,928)
3. *Revsunds tingslag*, 1649–1689 (101,020)
4. *Vendels socken*, 1615–45 (59,948)
5. *Vendels socken*, 1736–37 (55,780)
6. *Östra härads i Njudung*, 1602–1605 (34,956)

Documents related to the Church:

1. *Västerås recess*, 1527 (12,193)
2. *1571 års kyrkoordning* (49,043)
3. *Uppsala mötes beslut*, 1593 (26,969)
4. *1686 års kyrkolag* (25,799)

A gold standard of 40 randomly selected sentences from each text was compiled, i.e., in total 400 sentences. The gold standard was produced by manually annotating the sentences regarding verbs and complements. Because sentences are much longer in these texts than in contemporary texts, the 400 sentences together contain a total of 3,105 verbs. Each word form that was interpreted as a verb was annotated with the tag VB, and complements were enclosed in brackets labeled with their grammatical function. This is illustrated in Figure 1, which shows an annotated segment from the test corpus.

For comparison with contemporary text, we make use of a subset of the Stockholm-Umeå Corpus of contemporary Swedish text, SUC (Ejerhed and Källgren, 1997). This subset contains those segments in SUC that have been syntactically annotated and manually revised in the Swedish Treebank. In total, the subset includes approximately 20,000 tokens. Since the tagger used in the experiments on historical texts is trained on the whole of SUC, we had to slightly modify the extraction algorithm in order not to evaluate on the same data as the tagger has been trained. When testing the algorithm on contemporary text, we therefore trained a new model for the tagger, including all tokens in SUC except for the tokens reserved for evaluation.

Anklagadhes/VB$_1$	**Was accused/VB$_1$**
[SS$_{vb1}$	[SS$_{vb1}$
ryttaren	the horse-rider
Hindrik	Hindrik
Hyldh	Hyldh
SS$_{vb1}$]	SS$_{vb1}$]
hwilken	who
[OO$_{vb2}$	[OO$_{vb2}$
mökrenkningh	rape
OO$_{vb2}$]	OO$_{vb2}$]
giordt/VB$_2$	**done/VB$_2$**
medh	with
en	a
gienta	girl
Elin	Elin
Eriksdotter	Eriksdotter
i	in
Sikås	Sikås
,	,
hwarföre	why
rätten	the Court
honnom	him
tilspordhe/VB$_3$	**asked/VB$_3$**
[OO$_{vb3}$	[OO$_{vb3}$
om	if
han	he
[OO$_{vb4}$	[OO$_{vb4}$
dhetta	this
OO $_{vb4}$]	OO $_{vb4}$]
giordt/VB$_4$	**done/VB$_4$**
hafwer/VB$_5$	**has/VB$_5$**
OO $_{vb3}$]	OO $_{vb3}$]

Figure 1: Annotated segment in the test corpus.

4.2 Evaluation Metrics

In order to get a more fine-grained picture of the system's performance, we want to evaluate three different aspects:

1. Identification of verbs
2. Identification of complements
3. Identification of holistic verb constructions

The identification of verbs depends only on the part-of-speech tagger and can be evaluated using traditional precision and recall measures, comparing the tokens analysed as verbs by the tagger to the tokens analysed as verbs in the gold standard.

The identification of complements depends on both the tagger and the parser and can also be

evaluated using precision and recall measures. In this case, every complement identified by the parser is compared to the complements annotated in the gold standard. Precision is the number of correct complements found by the parser divided by the total number of complements output by the parser, while recall is the number of correct complements found by the parser divided by the number of complements in the gold standard. We do not take the labels into account when assessing complements as correct or incorrect. The motivation for this is that the overall aim of the complement extraction is to present verb expressions to historians, for them to consider whether they are describing work activities or not. In this context, only textual strings will be of interest, and grammatical function labels are ignored. For example, assume that the gold standard is:

lefverere [IO honom] [OO Sädh]
deliver [IO him] [OO grain]

and that the system produces:

lefverere [OO honom]
deliver [OO him]

In this context, the complement *honom* ("him") will be regarded as correct, even though it has been analysed as a direct object instead of an indirect object. On the other hand, the evaluation of complement identification is strict in that it requires the complement found to coincide exactly with the complement in the gold standard. For example, assume the gold standard is:

effterfrågat [OA om sinss manss dödh]
asked [OA about her husband's death]

and that the system produces:

effterfrågat [OA om sinss manss]
asked [OA about her husband's]

In this case, the complement will not be regarded as correct because it does not cover exactly the same textual string as the gold standard annotation.

The identification of holistic verb constructions, that is, a verb and all its complements, depends on the identification of verbs and complements, as well as the optional filtering of complements using valency dictionaries. Here we want to evaluate the entire text segment extracted in a way that is relevant for the intended application of the system. First of all, this means that partially correct constructions should be taken into account. Consider again the earlier example:

effterfrågat [OA om sinss manss dödh]
asked [OA about her husband's death]

and assume that the system produces:

effterfrågat [OA om sinss manss]
asked [OA about her husband's]

As noted above, this complement would be considered incorrect in the precision/recall evaluation of complement extraction, even though only one word is missing as compared to the gold standard, and the output would probably still be valuable to the end-user. Secondly, we should consider the total segment extracted for a verb including all complements, rather than inspecting each complement separately.

In order to reflect partially correct complements and take the total segment extracted for each verb into account, we use a string-based evaluation method for the identification of holistic verb constructions. In this evaluation, all labels and brackets are removed before comparing the segments extracted to the segments in the text corpus and each extracted instance is classified as falling into one of the four following categories:

- Fully correct complement set (F)
- Partially correct complement set (P)
- Incorrect complement set (I)
- Missing complement set (M)

A complement set is regarded as fully correct if the output string generated by the system is identical to the corresponding gold standard string. Since labels and brackets have been removed, these analyses will be regarded as identical:

lemnat [IO swaranden] [OO tid]
given [IO the defendant] [OO time]

lemnat [OO swaranden tid]
given [OO the defendant time]

A complement set is regarded as partially correct if the output string generated by the system has a non-empty overlap with the corresponding gold standard string. For example, the following three sets of analyses will be considered as partially correct (gold standard top, system output bottom):

lefverere [IO honom] [OO Sädh]
deliver [IO him] [OO Grain]
lefverere [OO honom]
deliver [OO him]

effterfrågat [OA om sinss manss dödh]
asked [OA about her husband's death]
effterfrågat [OA om sinss manss]
asked [OA about her husband's]

betale [PL åter] [IO här Mattz] [OO Rågen]
pay [PL back] [IO mister Mattz] [OO the Rye]
betale [OO åter här Mattz]
pay [OO back mister Mattz]

A (non-empty) complement set is regarded as incorrect if the output string has no overlap with the gold standard string. Finally, a complement set is regarded as missing if the output string is empty but the gold standard string is not. It is worth noting that the four categories are mutually exclusive.

5 Results and Discussion

In this section, we evaluate the identification of verbs, complements and holistic verb constructions using the data and metrics described in the previous section.

5.1 Verbs

Results on the identification of verbs using part-of-speech tagging, with and without normalisation, are reported in Table 2. As can be seen, recall drastically increases when normalisation rules are applied prior to tagging, even though the normalisation rules used in this experiment are formulated based on a subset of one single 17th century text, and the test corpus contains samples of various text types ranging from 1527–1737. Normalisation also has a small positive effect on precision, and the best result for historical texts is 78.4% precision and 76.0% recall. This is slightly lower than the results presented by Pettersson and Nivre (2011) where only 17th century text was used for evaluation, indicating that the normalisation rules are somewhat biased towards 17th century text, and that the results could be improved if normalisation were adapted to specific time periods. It is also worth noting that the results are substantially lower for historical text than the results for contemporary text, with precision and recall at 99.1%, but still high enough to be useful in the intended context of application.

	Precision	**Recall**	**F-score**
Raw	75.4	60.0	66.9
Norm	78.4	76.0	77.2
SUC	99.1	99.1	99.1

Table 2: Identification of verbs by tagging. Raw = Unnormalised input text. Norm = Normalisation of input prior to tagging. SUC = Subset of Stockholm-Umeå corpus of contemporary Swedish texts, as described in section 4.1.

Tokens that are still erroneously analysed by the tagger include the following cases:

- tokens where the old spelling is identical to an existing, but different, word form in contemporary language; for example, the spelling *skal* would in contemporary language be considered a noun ("shell") but in the old texts this spelling is used for a word that is nowadays spelled *skall* ("shall/should") and should be regarded as a verb;

- ambiguous words; for example, past participles are often spelled the same way as the corresponding past tense verb, but participles are not regarded as verb forms in our experiments;[4]

- tokens that have not been normalised enough and thus do not correspond to a word form recognised by the tagger, e.g., the word form *lemnas* which in contemporary language should be spelled as *lämnas* ("be left").

[4]Participles are only used adjectivally in Swedish, as the perfect tense is formed using a distinct supine form of the verb.

	Precision	Recall	F-score
Raw	24.8	27.5	26.1
Norm	28.3	28.2	28.2
+Valency	33.1	25.5	28.8
SUC	68.2	70.7	69.5
+Valency	71.8	56.2	63.0

Table 3: Identification of complements by parsing. Raw = Unnormalised input text. Norm = Normalisation of input prior to tagging and parsing. +Valency = Adding valency filtering to the setting in the preceding row. SUC = Subset of Stockholm-Umeå corpus of contemporary Swedish texts, as described in section 4.1.

	F	P	I	M
Raw	32.6	20.3	29.3	17.8
Norm	34.5	19.5	25.2	20.8
+Valency	38.7	16.9	18.9	25.5
SUC	30.3	54.2	9.1	6.4
+Valency	30.8	47.9	6.8	14.6

Table 4: Identification of holistic verb constructions. F = Fully correct, P = Partially correct, I = Incorrect, M = Missing. Raw = Unnormalised input text. Norm = Normalisation of input prior to tagging and parsing. +Valency = Adding valency filtering to the setting in the preceding row. SUC = Subset of Stockholm-Umeå corpus of contemporary Swedish texts, as described in section 4.1.

5.2 Complements

Recall and precision for the identification of complements using parsing are presented in Table 3. In this case, normalisation has a smaller effect than in the case of tagging and affects precision more than recall. Adding a filter that eliminates unlikely complements based on the valency frame of the verb in existing dictionaries predictably improves precision at the expense of recall and results in a small F-score improvement.

Again, the best scores on the historical texts are much lower than the corresponding results for contemporary text, with an F-score of 28.8% in the former case and 69.5% in the latter, but it is worth remembering that precision and recall on exactly matching complements is a harsh metric that is not directly relevant for the intended application. Finally, it is worth noting that the valency filter has a large negative impact on recall for the modern texts, resulting in a decrease in the F-score, which indicates that the parser in this case is quite successful at identifying complements (in the wide sense) that are not covered by the valency dictionaries.

5.3 Verb Constructions

As argued in section 4.2, precision and recall measures are not sufficient for evaluating the extraction of holistic verb constructions. A more relevant assessment is made by counting the number of *fully correct*, *partially correct*, *incorrect* and *missing* complement sets for the verbs identified. Table 4 summarises the results in accordance with this metric.

First of all, we see that normalisation again has a rather small effect on overall results, increasing the number of fully correct constructions and decreasing the number of incorrect constructions, but also leading to an increase in the number of missing complements. Adding the valency filter to remove unlikely complements has a similar effect and increases the percentage of correctly extracted verb constructions to 38.7% while decreasing the share of incorrect constructions to 18.9%. However, it also increases the percentage of verbs with missing complement sets from 20.8% to 25.5%. This is partly due to the fact that some of the verbs are used in a slightly different way in historical text as compared to contemporary text, meaning that the valency frames are not as reliable. For example, the verb *avstå* ("refrain") in the historical corpus is used with a direct object, as in *Anders Andersson afstådt sitt skatte hemman* ("Anders Andersson refrained his homestead"), whereas in a contemporary context this verb would more likely be used with a prepositional complement, *avstå* **från** *någonting* ("refrain **from** something").

In total, 55.6% of the verbs are assigned a fully or partially correct set of complements. This is again lower than the result for contemporary texts (78.7%), but the difference is smaller than for the previous metrics, which is encouraging given that the evaluation in this section is most relevant for the intended application. Moreover, it is worth noting that the difference is mostly to be found in the category of partially correct constructions, where the best result for modern texts is 54.2%, to be compared to 16.9% for the historical texts. With respect to fully correct constructions, however, the results are actually better for the histor-

ical texts than for the modern texts, 38.7% vs. 30.8%, a rather surprising positive result.

6 Conclusion

We have presented a method for automatically extracting verbs and their complements from historical Swedish texts, more precisely texts from the Early Modern era (1550–1800), with the aim of providing language technology support for historical research. We have shown that it is possible to use existing contemporary NLP tools and dictionaries for this purpose, provided that the input text is first (automatically) normalised to a more modern spelling. With the best configuration of our tools, we can identify verbs with an F-score of 77.2% and find a partially or completely correct set of complements for 55.6% of the verbs. To the best of our knowledge, these are the first results of their kind.

In addition to presenting a method for the identification of verb constructions, we have also proposed a new evaluation framework for such methods in the context of information extraction for historical research. As a complement to standard precision and recall metrics for verbs and their complements, we have evaluated the text segments extracted using the categories *fully correct*, *partially correct*, *incorrect*, and *missing*. One important topic for future research is to validate this evaluation framework by correlating it to the perceived usefulness of the system when used by historians working on the Gender and Work Database. Preliminary experiments using a prototype system indicate that this kind of support can in fact reduce the time-consuming, manual work that is currently carried out by historians and other researchers working with older texts.

Another topic for future research concerns the variation in performance across time periods and text types. In the current evaluation, court records and papers related to the Church ranging from 1527 to 1737 have been sampled in the gold standard. It would be interesting to explore in more detail how the program performs on the oldest texts as compared to the youngest texts, and on court records as compared to the other genres.

References

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

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

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

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

Eva Ejerhed and Gunnel Källgren. 1997. Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

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

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

Beáta B. Megyesi. 2009. The open source tagger HunPos for Swedish. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA)*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. 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, Genoa, Italy, May.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006b. 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, Genoa, Italy, May.

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

Marco Pennacchiotti and Fabio Massimo Zanzotto. 2008. Natural language processing across time: An empirical investigation on Italian. In *Advances in Natural Language Processing. GoTAL, LNAI*, volume 5221, pages 371–382.

Eva Pettersson and Joakim Nivre. 2011. 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*, pages 87–95, Portland, OR,

USA, June. Association for Computational Linguistics.

Vito Rocio, Mário Amado Alves, José Gabriel Lopes, Maria Francisca Xavier, and Graça Vicente. 1999. Automated creation of a partially syntactically annotated corpus of Medieval Portuguese using contemporary Portuguese resources. In *Proceedings of the ATALA workshop on Treebanks*, Paris, France.

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