Automatic Verb Extraction from Historical Swedish Texts

Eva Pettersson
Department of Linguistics and Philology
Uppsala University
Swedish National Graduate School
of Language Technology
eva.pettersson@lingfil.uu.se

Joakim Nivre
Department of Linguistics and Philology
Uppsala University
joakim.nivre@lingfil.uu.se

Abstract

Even though historical texts reveal a lot of interesting information on culture and social structure in the past, information access is limited and in most cases the only way to find the information you are looking for is to manually go through large volumes of text, searching for interesting text segments. In this paper we will explore the idea of facilitating this time-consuming manual effort, using existing natural language processing techniques. Attention is focused on automatically identifying verbs in early modern Swedish texts (1550–1800). The results indicate that it is possible to identify linguistic categories such as verbs in texts from this period with a high level of precision and recall, using morphological tools developed for present-day Swedish, if the text is normalised into a more modern spelling before the morphological tools are applied.

1 Introduction

Historical texts constitute a rich source of data for researchers interested in for example culture and social structure over time. It is however a very time-consuming task to manually search for relevant passages in the texts available. It is likely that language technology could substantially reduce the manual effort involved and thus the time needed to access this information, by automatically suggesting sections that may be of interest to the task at hand. The interesting text segments could be identified using for example semantic features or morphological and syntactic cues in the text.

This would however require natural language processing tools capable of handling historical texts, which are in many respects different from contemporary written language, concerning both spelling and syntax. Ideally, one would of course like to have tools developed specifically for the time period of interest, and emerging efforts to develop resources and tools for historical languages are therefore welcome. Despite these efforts, however, it is unlikely that we will have anything close to complete coverage of different time periods even for a single language within the foreseeable future.

In this paper, we will therefore instead examine the possibility of improving information access in historical texts by adapting language technology tools developed for contemporary written language. The work has been carried out in close cooperation with historians who are interested in what men and women did for a living in the early modern Swedish society (1550–1800). We will hence focus on identifying linguistic categories in Swedish texts from this period. The encouraging results show that you may successfully analyse historical texts using NLP tools developed for contemporary language, if analysis is preceded by an orthographic normalisation step.

Section 2 presents related work and characteristics of historical Swedish texts. The extraction method is defined in section 3. In section 4 the experiments are described, while the results are presented in section 5. Section 6 describes how the verb extraction tool is used in ongoing historical research. Finally, conclusions are drawn in section 7.

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 87–95,
Portland, OR, USA, 24 June 2011. ©2011 Association for Computational Linguistics

2 Background

2.1 Related Work

There are still not many studies performed on natural language processing of historical texts. Pennacchiotti and Zanzotto (2008) used contemporary dictionaries and analysis tools to analyse Italian texts from the period 1200–1881. The results showed that 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. The morphological analyser used in the study reached an accuracy of 0.48 (as compared to 0.91 for modern text), while the part-of-speech tagger yielded an accuracy of 0.54 (as compared to 0.97 for modern text).

Rocio et al. (1999) used a grammar of contemporary Portuguese to syntactically annotate medieval Portuguese texts. To adapt the parser to the medieval language, a lexical analyser was added including a dictionary and inflectional rules for medieval Portuguese. This combination proved to be successful for partial parsing of medieval Portuguese texts, even though there were some problems with grammar limitations, dictionary incompleteness and insufficient part-of-speech tagging.

Oravecz et al. (2010) tried a semi-automatic approach to create an annotated corpus of texts from the Old Hungarian period. The annotation was performed in three steps: 1) sentence segmentation and tokenisation, 2) standardisation/normalisation, and 3) morphological analysis and disambiguation. They concluded that normalisation is of vital importance to the performance of the morphological analyser.

For the Swedish language, Borin et al. (2007) proposed a named-entity recognition system adapted to Swedish literature from the 19th century. The system recognises Person Names, Locations, Organisations, Artifacts (food/wine products, vehicles etc), Work&Art (names of novels, sculptures etc), Events (religious, cultural etc), Measure/Numerical expressions and Temporal expressions. The named entity recognition system was evaluated on texts from the Swedish Literature Bank without any adaptation, showing problems with spelling variation, inflectional differences, unknown names and structural issues (such as hyphens splitting a single name into several entities).[1] Normalising the texts before applying the named entity recognition system made the f-score figures increase from 78.1% to 89.5%.

All the results presented in this section indicate that existing natural language processing tools are not applicable to historical texts without adaptation of the tools, or the source text.

2.2 Characteristics of Historical Swedish Texts

Texts from the early modern Swedish period (1550–1800) differ from present-day Swedish texts both concerning orthography and syntax. Inflectional differences include a richer verb paradigm in historical texts as compared to contemporary Swedish. The Swedish language was also strongly influenced by other languages. Evidence of this is the placement of the finite verb at the end of relative clauses in a German-like fashion not usually found in Swedish texts, as in *...om man i hächtelse sitter* as compared to *om man sitter i häkte* (*"...if you in custody are"* vs *"...if you are in custody"*).

Examples of the various orthographic differences are the duplication of long vowels in words such as *saak* (*sak* "thing") and *stoor* (*stor* "big/large"), the use of of *fv* instead of *v*, as in *öfver* (*över* "over"), and *gh* and *dh* instead of the present-day *g* and *d*, as in *någhon* (*någon* "somebody") and *fadhren* (*fadern* "the father") (Bergman, 1995).

Furthermore, the lack of spelling conventions causes the spelling to vary highly between different writers and text genres, and even within the same text. There is also great language variation in texts from different parts of the period.

3 Verb Extraction

In the following we will focus on identifying verbs in historical Swedish texts from the period 1550–1800. The study has been carried out in cooperation with historians who are interested in finding out what men and women did for a living in the early modern Swedish society. One way to do this would be to search for occupational titles occurring in the text. This is however not sufficient since many people, especially women, had no occupational title. Occupational titles are also vague, and may include several subcategories of work. In the material

[1] http://litteraturbanken.se/

already (manually) analysed by the historians, occupation is often described as a verb with a direct object. Hence, automatically extracting and displaying the verbs in a text could help the historians in the process of finding relevant text segments. The verb extraction process developed for this purpose is performed in maximally five steps, as illustrated in figure 1.

The first step is tokenisation. Each token is then optionally matched against dictionaries covering historical Swedish. Words not found in the historical dictionaries are normalised to a more modern spelling before being processed by the morphological analyser. Finally, the tagger disambiguates words with several interpretations, yielding a list of all the verb candidates in the text. In the experiments, we will examine what steps are essential, and how they are combined to yield the best results.

3.1 Tokenisation

Tokenisation is performed using an in-house standard tokeniser. The result of the tokenisation is a text segmented into one token per line, with a blank line marking the start of a new sentence.

3.2 Historical Dictionaries

After tokenisation, the tokens are optionally matched against two historical dictionaries distributed by *The Swedish Language Bank*:[2]

- **The Medieval Lexical Database**
  A dictionary describing Medieval Swedish, containing approximately 54 000 entries from the following three books:
  - K.F. Söderwalls *Ordbok Öfver svenska medeltids-språket, vol I-III* (Söderwall, 1918)
  - K.F. Söderwalls *Ordbok Öfver svenska medeltids-språket, vol IV-V* (Söderwall, 1973)
  - C.J. Schlyters *Ordbok till Samlingen af Sweriges Gamla Lagar* (Schlyter, 1877)
- **Dalin's Dictionary**
  A dictionary covering 19th Century Swedish, created from the printed version of *Ordbok Öfver svenska språket, vol I–II* by Dalin (1855). The dictionary contains approximately 64 000 entries.

The dictionaries cover medieval Swedish and 19th century Swedish respectively. We are actually interested in the time period in between these two periods, but it is assumed that these dictionaries are close enough to cover words found in the early modern period as well. It should further be noticed that the electronically available versions of the dictionaries are still in an early stage of development. This means that coverage varies between different word classes, and verbs are not covered to the same extent as for example nouns. Words with an irregular inflection (which is often the case for frequently occurring verbs) also pose a problem in the current dictionaries.

3.3 Normalisation Rules

Since both the morphological analyser and the tagger used in the experiments are developed for handling modern Swedish written language, running a text with the old Swedish spelling preserved presumably means that these tools will fail to assign correct analyses in many cases. Therefore, the text is optionally transformed into a more modern spelling, before running the document through the analysis tools.

The normalisation procedure differs slightly for morphological analysis as compared to tagging. There are mainly two reasons why the same set of normalisation rules may not be optimally used both for the morphological analyser and for the tagger. First, since the tagger (unlike the morphological analyser) is context sensitive, the normalisation rules developed for the tagger need to be designed to also normalise words surrounding verbs, such as nouns, determiners, etc. For the morphological analyser, the main focus in formulating the rules has been on handling verb forms. Secondly, to avoid being limited to a small set of rules, an incremental normalisation procedure has been used for the morphological analyser in order to maximise recall without sacrificing precision. In this incremental process, normalisation rules are applied one by one, and the less confident rules are only applied to words not identified by the morphological analyser in the previous

[2] http://spraakbanken.gu.se/

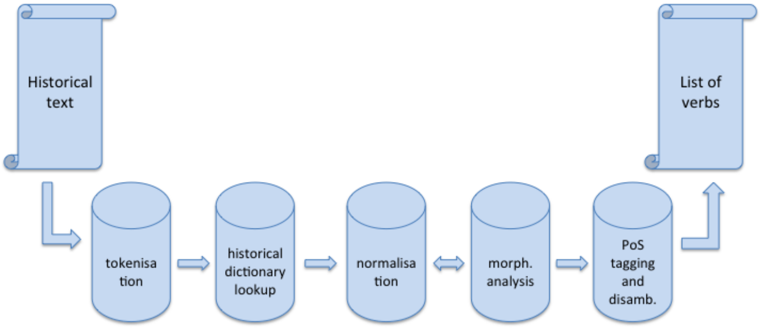


Figure 1: Overview of the verb extraction experiment

normalisation step. The tagger on the other hand is robust, always yielding a tag for each token, even in cases where the word form is not present in the dictionary. Thus, the idea of running the normalisation rules in an incremental manner is not an option for the tagger.

The total set of normalisation rules used for the morphological analyser is 39 rules, while 29 rules were defined for the tagger. The rules are inspired by (but not limited to) some of the changes in the reformed Swedish spelling introduced in 1906 (Bergman, 1995). As a complement to the rules based on the spelling reform, a number of empirically designed rules were formulated, based on the development corpus described in section 4.1. The empirical rules include the rewriting of verbal endings (e.g. *begärade* – *begärde* "requested" and *utviste* – *utvisade* "deported"), transforming double consonants into a single consonant (*vetta* – *veta* "know", *prövass* – *prövas* "be tried") and vice versa (*upsteg* – *uppsteg* "rose/ascended", *viste* – *visste* "knew").

3.4 Morphological Analysis and Tagging

SALDO is an electronically available lexical resource developed for present-day written Swedish. It is based on *Svenskt AssociationsLexikon* (SAL), a semantic dictionary compiled by Lönngren (1992). The first version of the SALDO dictionary was released in 2008 and comprises 72 396 lexemes. Inflectional information conforms to the definitions in Nationalencyklopedins ordbok (1995), Svenska Akademiens ordlista över svenska språket (2006) and Svenska Akademiens grammatik (1999). Apart from single word entries, the SALDO dictionary also contains approximately 2 000 multi-word units, including 1 100 verbs, mainly particle verbs (Borin et al., 2008). In the experiments we will use SALDO version 2.0, released in 2010 with a number of words added, resulting in a dictionary comprising approximately 100 000 entries.

When running the SALDO morphological analyser alone, a token is always considered to be a verb if there is a verb interpretation present in the dictionary, regardless of context. For example, the word *för* will always be analysed both as a verb (*bring*) and as a preposition (*for*), even though in most cases the prepositional interpretation is the correct one.

When running the maximum five steps in the verb extraction procedure, the tagger will disambiguate in cases where the morphological analyser has produced both a verb interpretation and a non-verb interpretation. The tagger used in this study is HunPOS (Halácsy et al., 2007), a free and open source reimplementation of the HMM-based TnT-tagger by Brants (2000). Megyesi (2008) showed that the HunPOS tagger trained on the Stockholm-Umeå Corpus (Gustafson-Capková and Hartmann, 2006) is one of the best performing taggers for Swedish texts.

4 Experiments

This section describes the experimental setup including data preparation and experiments.

4.1 Data Preparation

A subset of *Per Larssons dombok*, a selection of court records from 1638, was used as a basis for developing the automatic verb extraction tool. This text consists of 11 439 tokens in total, and was printed by Edling (1937). The initial 984 tokens of the text were used as development data, i.e. words used when formulating the normalisation rules, whereas the rest of the text was used solely for evaluation.

A gold standard for evaluation was created, by manually annotating all the verbs in the text. For the verb annotation to be as accurate as possible, the same text was annotated by two persons independently, and the results analysed and compared until consensus was reached. The resulting gold standard includes 2 093 verbs in total.

4.2 Experiment 1: Normalisation Rules

In the first experiment we will compare morphological analysis results before and after applying normalisation rules. To investigate what results could optimally be expected from the morphological analysis, SALDO was also run on present-day Swedish text, i.e. the Stockholm-Umeå Corpus (SUC). SUC is a balanced corpus consisting of a number of different text types representative of the Swedish language in the 1990s. The corpus consists of approximately one million tokens, distributed among 500 texts with approximately 2 000 tokens in each text. Each word in the corpus is manually annotated with part of speech, lemma and a number of morphological features (Gustafson-Capková and Hartmann, 2006).

4.3 Experiment 2: Morphological Analysis and Tagging

In the second experiment we will focus on the combination of morphological analysis and tagging, based on the following settings:

morph A token is always considered to be a verb if the morphological analysis contains a verb interpretation.

tag A token is always considered to be a verb if it has been analysed as a verb by the tagger.

morph *or* **tag** A token is considered to be a verb if there is a morphological verb analysis **or** if it has been analysed as a verb by the tagger.

morph *and* **tag** A token is considered to be a verb if there is a morphological verb analysis **and** it has been tagged as a verb.

To further refine the combination of morphological analysis and tagging, a more fine-grained disambiguation method was introduced, where the tagger is only used in contexts where the morphological analyser has failed to provide an unambiguous interpretation:

morph + tag A token is considered to be a verb if it has been unambiguously analysed as a verb by SALDO. Likewise a token is considered not to be a verb, if it has been given one or more analyses from SALDO, where none of the analyses is a verb interpretation. If the token has been given both a verb analysis and a non-verb analysis by SALDO, the tagger gets to decide. The tagger also decides for words not found in SALDO.

4.4 Experiment 3: Historical Dictionaries

In the third experiment, the historical dictionaries are added, using the following combinations:

medieval A token is considered to be a verb if it has been unambiguously analysed as a verb by the medieval dictionary. Likewise a token is considered not to be a verb, if it has been given one or more analyses from the medieval dictionary, where none of the analyses is a verb interpretation. If the token has been given both a verb analysis and a non-verb analysis by the medieval dictionary, or if the token is not found in the dictionary, the token is processed by the morphological analyser and the tagger as described in setting *morph + tag*.

19c A token is considered to be a verb if it has been unambiguously analysed as a verb by the 19th century dictionary. Likewise a token is considered not to be a verb, if it has been given one or more analyses from the 19th century dictionary, where none of the analyses is a verb interpretation. If the token has been given both a verb analysis and a non-verb analysis by the 19th century dictionary, or if the token is not found in the dictionary, the token is processed by the morphological analyser and the tagger as described in setting *morph + tag*.

medieval + 19c A token is considered to be a verb if it has been unambiguously analysed as a verb by the medieval dictionary. Likewise a token is considered not to be a verb, if it has been given one or more analyses from the medieval dictionary, where none of the analyses is a verb interpretation. If the token has been given both a verb analysis and a non-verb analysis by the medieval dictionary, or if the token is not found in the dictionary, the token is matched against the 19th century dictionary before being processed by the morphological analyser and the tagger as described in setting *morph + tag*.

19c + medieval A token is considered to be a verb if it has been unambiguously analysed as a verb by the 19th century dictionary. Likewise a token is considered not to be a verb, if it has been given one or more analyses from the 19th century dictionary, where none of the analyses is a verb interpretation. If the token has been given both a verb analysis and a non-verb analysis by the 19th century dictionary, or if the token is not found in the dictionary, the token is matched against the medieval dictionary before being processed by the morphological analyser and the tagger as described in setting *morph + tag*.

5 Results

5.1 Normalisation Rules

Running the SALDO morphological analyser on the test text with the old Swedish spelling preserved, meant that only 30% of the words were analysed at all. Applying the normalisation rules before the morphological analysis is performed, drastically increases recall. After only 5 rules have been applied, recall is increased by 11 percentage units, and adding another 5 rules increases recall by another 26 percentage units. All in all, recall increases from 30% for unnormalised text to 83% after all normalisation rules have been applied, whereas precision increases from 54% to 66%, as illustrated in table 1.

Recall is still significantly higher for contemporary Swedish texts than for the historical text (99% as compared to 83% with the best normalisation settings). Nevertheless, the rapid increase in recall when applying the normalisation rules is very promising, and it is yet to be explored how good results it is possible to reach if including more normalisation rules.

	Precision	**Recall**	**f-score**
raw data	0.54	0.30	0.39
5 rules	0.61	0.41	0.49
10 rules	0.66	0.67	0.66
15 rules	0.66	0.68	0.67
20 rules	0.67	0.73	0.70
25 rules	0.66	0.78	0.72
30 rules	0.66	0.79	0.72
35 rules	0.66	0.82	0.73
39 rules	0.66	0.83	0.74
SUC corpus	0.53	0.99	0.69

Table 1: Morphological analysis results using SALDO version 2.0, before and after incremental application of normalisation rules, and compared to the Stockholm-Umeå corpus of contemporary Swedish written language.

5.2 Morphological Analysis and Tagging

Table 2 presents the results of combining the SALDO morphological analyser and the HunPOS tagger, using the settings described in section 4.3.

	Precision	**Recall**	**f-score**
morph	0.66	0.83	0.74
tag	0.81	0.86	0.83
morph *or* **tag**	0.61	0.92	0.74
morph *and* **tag**	0.92	0.80	0.85
morph + tag	0.82	0.88	0.85

Table 2: Results for normalised text, combining morphological analysis and tagging. morph = morphological analysis using SALDO. tag = tagging using HunPOS.

As could be expected, the tagger yields higher precision than the morphological anlayser, due to the fact that the morphological analyser renders all analyses for a word form given in the dictionary, regardless of context. The results of combining the morphological analyser and the tagger are also quite expected. In the case where a token is considered to be a verb if there is a morphological verb analysis *or* it has been analysed as a verb by the tagger, a very high level of recall (92%) is achieved at the expense of low precision, whereas the opposite is true for the case where a token is considered to be a verb if there is a morphological verb analysis *and* it has been tagged as a verb. Using the tagger for disambiguation only in ambiguous cases yields the best results. It should be noted that using the morph-and-tag setting results in the same f-score as the disambiguation setting. However, the disambiguation setting performs better in terms of recall, which is of importance to the historians in the project at hand. Another advantage of using the disambiguation setting is that the difference between precision and recall is less.

5.3 Historical Dictionaries

The results of using the historical dictionaries are presented in table 3.

	Precision	**Recall**	**f-score**
morph + tag	0.82	0.88	0.85
medieval	0.82	0.81	0.81
19c	0.82	0.86	0.84
medieval + 19c	0.81	0.79	0.80
19c + medieval	0.81	0.79	0.80

Table 3: Results for normalised text, combining historical dictionaries and contemporary analysis tools. medieval = *Medieval Lexical Database*. 19c = *Dalin's Dictionary*. morph = morphological analysis using SALDO. tag = tagging using HunPOS.

Adding the historical dictionaries did not improve the verb analysis results; actually the opposite is true. Studying the results of the analyses from the medieval dictionary, one may notice that only two verb analyses have been found when applied to the test text, and both of them are erroneous in this context (in both cases the word *lass* "load" as in the phrase *6 lass höö* "6 loads of hay"). Furthermore, the medieval dictionary produces quite a lot of non-verb analyses for commonly occurring verbs, for example *skola* (noun: "shool", verb: "should/shall"), *kunna* ("can/could"), *kom* ("come"), *finna* ("find") and *vara* (noun: "goods", verb: "be"). Another reason for the less encouraging results seems to be that most of the words actually found and analysed correctly are words that are correctly analysed by the contemporary tools as well, such as *i* ("in"), *man* ("man/you"), *sin* ("his/her/its"), *honom* ("him") and *in* ("into").

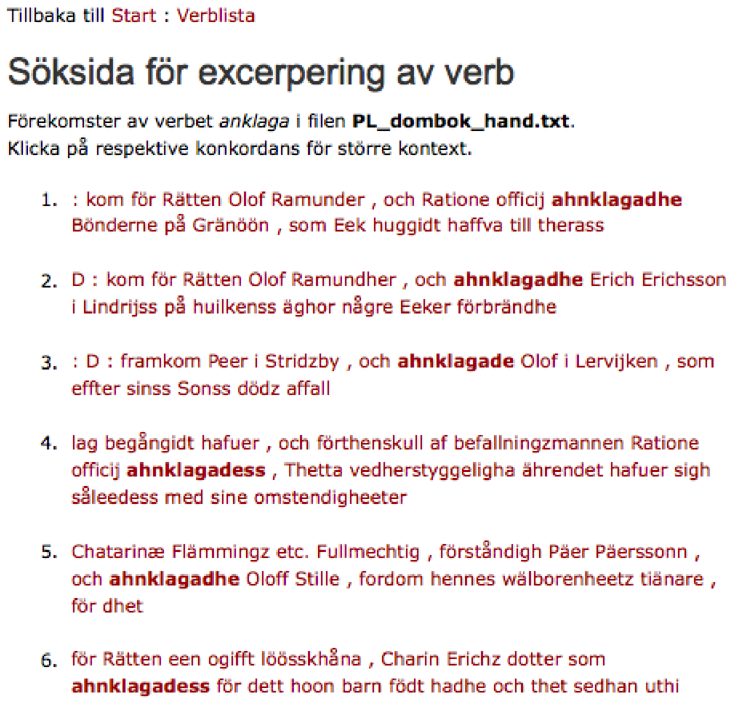
As for the 19th century dictionary, the same problems apply. For example, a number of frequent verb forms are analysed as non-verbs (e.g. *skall* "should/shall" and *ligger* "lies"). There are also non-verbs repeatedly analysed as verbs, such as *stadgar* ("regulations") and *egne* ("own"). As was the case for the medieval dictionary, most of the words analysed correctly by the 19th century dictionary are commonly occuring words that would have been correctly analysed by the morphological analyser and/or the tagger as well, for example *och* ("and"), *men* ("but") and *när* ("when").

6 Support for Historical Research

In the ongoing *Gender and Work* project at the Department of History, Uppsala University, historians are interested in what men and women did for a living in the early modern Swedish Society.[3] Information on this is registered and made available for research in a database, most often in the form of a verb and its object(s). The automatic verb extraction tool was developed in close cooperation with the Gender and Work participants, with the aim of reducing the manual effort involved in finding the relevant information to enter into the database.

The verb extraction tool was integrated in a prototypical graphical user interface, enabling the historians to run the system on historical texts of their choice. The interface provides facilities for uploading files, generating a list of all the verbs in the file, displaying verb concordances for interesting verbs, and displaying the verb in a larger context. Figure 2 illustrates the graphical user interface, displaying concordances for the verb *anklaga* ("accuse"). The historians found the interface useful and are interested in integrating the tool in the Gender and Work database. Further development of the verb extraction tool is now partly funded by the Gender and Work project.

[3] http://gaw.hist.uu.se/



Tillbaka till Start : Verblista

Söksida för excerpering av verb

Förekomster av verbet *anklaga* i filen **PL_dombok_hand.txt**.
Klicka på respektive konkordans för större kontext.

1. : kom för Rätten Olof Ramunder , och Ratione officij **ahnklagadhe** Bönderne på Gränöön , som Eek huggidt haffva till therass
2. D : kom för Rätten Olof Ramundher , och **ahnklagadhe** Erich Erichsson i Lindrijss på huilkenss äghor några Eeker förbrändhe
3. : D : framkom Peer i Stridzby , och **ahnklagade** Olof i Lervijken , som effter sinss Sonss dödz affall
4. lag begångidt hafuer , och förthenskull af befallningzmannen Ratione officij **ahnklagadess** , Thetta vedherstyggeligha ährendet hafuer sigh såleedess med sine omstendigheeter
5. Chatarinæ Flämmingz etc. Fullmechtig , förståndigh Päer Päerssonn , och **ahnklagadhe** Oloff Stille , fordom hennes wälborenheetz tiänare , för dhet
6. för Rätten een ogifft löösskhåna , Charin Erichz dotter som **ahnklagadess** för dett hoon barn födt hadhe och thet sedhan uthi

Figure 2: Concordances displayed for the verb *anklaga* (“accuse”) in the graphical user interface.

7 Conclusion

Today historians and other researchers working on older texts have to manually go through large volumes of text when searching for information on for example culture or social structure in historical times. In this paper we have shown that this time-consuming manual effort could be significantly reduced using contemporary natural language processing tools to display only those text segments that may be of interest to the researcher. We have described the development of a tool that automatically identifies verbs in historical Swedish texts using morphological analysis and tagging, and a prototypical graphical user interface, integrating this tool. The results indicate that it is possible to retrieve verbs in Swedish texts from the 17th century with 82% precision and 88% recall, using morphological tools for contemporary Swedish, if the text is normalised into a more modern spelling before the morphological tools are applied (recall may be increased to 92% if a lower precision is accepted).

Adding electronically available dictionaries covering medieval Swedish and 19th century Swedish respectively to the verb extraction tool, did not improve the results as compared to using only contemporary NLP tools. This seems to be partly due to the dictionaries still being in an early stage of development, where lexical coverage is unevenly spread among different word classes, and frequent, irregularly inflected word forms are not covered. It would therefore be interesting to study the results of the historical dictionary lookup, when the dictionaries are more mature.

Since the present extraction tool has been evaluated on one single text, it would also be interesting to explore how these extraction methods should be adapted to handle language variation in texts from different genres and time periods. Due to the lack of spelling conventions, it would also be interesting to see how the extraction process performs on texts from the same period and genre, but written by different authors. Future work also includes experiments on identifying linguistic categories other than verbs.

References

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

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. *SALDO 1.0 (Svenskt associationslexikon version 2).* Språkbanken, University of Gothenburg.

Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 2007. *Naming the Past: Named Entity and Anomacy Recognition in 19th Century Swedish Literature).* In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 1–8. Prague, Czech Republic.

Bra Böcker. 1995. *Nationalencyklopedins ordbok.* Bra Böcker, Höganäs.

Thorsten Brants. 2000. *TnT - A Statistical Part-of-Speech Tagger.* In: Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-00), Seattle, Washington, USA.

Anders Fredrik Dalin. 1850–1855. *Ordbok Öfver svenska språket. Vol I–II.* Stockholm.

Nils Edling. 1937. *Uppländska domböcker. jämte inledning, förklaringar och register utgivna genom Nils Edling.* Uppsala.

Sofia Gustafson-Capková and Britt Hartmann. December 2006. *Manual of the Stockholm Umeå Corpus version 2.0.* Description of the content of the SUC 2.0 distribution, including the unfinished documentation by Gunnel Källgren.

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 Companion Volume Proceedings of the Demo and Poster Sessions, pages 209–212. Association for Computational Linguistics, Prague, Czech Republic.

Lennart Lönngren. 1992. *Svenskt associationslexikon, del I–IV.* Department of Linguistics and Philology, Uppsala University.

Beáta B. Megyesi. 2008. *The Open Source Tagger HunPoS for Swedish.* Department of Linguistics and Philology, Uppsala University.

Csaba Oravecz, Bálint Sass, and Eszter Simon 2010. *Semi-automatic Normalization of Old Hungarian Codices.* In: Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010). Pages 55–59. 16 August, 2010 Faculty of Science, University of Lisbon Lisbon, Portugal.

Marco Pennacchiotti and Fabio Massimo Zanzotto 2008. *Natural Language Processing Across Time: An Empirical Investigation on Italian.* In: Aarne Ranta and Bengt Nordström (Eds.): Advances in Natural Language Processing. GoTAL 2008, LNAI Volume 5221, pages 371–382. Springer-Verlag Berlin Heidelberg.

Vitor 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.

Carl Johan Schlyter. 1877. *Ordbok till Samlingen af Sweriges Gamla Lagar.* Lund.

Svenska Akademien. 2006. *Svenska Akademiens ordlista över svenska språket.* Norstedts Akademiska Förlag, Stockholm.

Knut Fredrik Söderwall. 1884–1918. *Ordbok Öfver svenska medeltids-språket, vol I–III.* Lund.

Knut Fredrik Söderwall. 1953–1973. *Ordbok Öfver svenska medeltids-språket, vol IV–V.* Lund.

Ulf Teleman, Staffan Hellberg, and Erik Andersson. 1999. *Svenska Akademiens grammatik.* Norstedts Ordbok, Stockholm.