HistSearch – Implementation and Evaluation of a Web-based Tool for Automatic Information Extraction from Historical Text

Eva Pettersson[1], Jonas Lindström[2], Benny Jacobsson[2], Rosemarie Fiebranz[2]

[1] Department of Linguistics and Philology, Uppsala University
[2] Department of History, Uppsala University

eva.pettersson@lingfil.uu.se, jonas.lindstrom@hist.uu.se, benny.jacobsson@hist.uu.se, rosemarie.fiebranz@hist.uu.se

Abstract

Due to a lack of NLP tools adapted to the task of analysing historical text, historians and other researchers in humanities often need to manually search through large volumes of text in order to find certain pieces of information of interest to their research. In this paper, we present a web-based tool for automatic information extraction from historical text, with the aim of facilitating this time-consuming process. We describe 1) the underlying architecture of the system, based on spelling normalisation succeeded by tagging and parsing using tools available for the modern language, 2) a prototypical graphical user interface used by the historians, and 3) a thorough manual evaluation of the tool performed by the actual users, i.e. the historians, when applied to the specific task of extracting and presenting verb phrases describing work in Early Modern Swedish text. The main contribution is the manual evaluation, which takes both quantitative and qualitative aspects into account, and is compared to automatic evaluation results. We show that spelling normalisation is successful for the task of tagging and lemmatisation, meaning that the words analysed as verbs by the tool are mostly considered as verbs by the historians as well. We also point out the further work needed for improving parsing and ranking performance, in order to make the tool really useful in the extraction process.

M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch (eds.): Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016, published at http://ceur-ws.org

1 Introduction

Historical text is a rich source of information for historians, philologists and other researchers in humanities. Nevertheless, a considerable amount of historical source material remains unexplored. One reason for this is that a large proportion of historical text is still only available in hand-written or possibly printed form, limiting the possibilities to access the information concealed in these texts. Furthermore, even for properly digitised text, there is a lack of language technology tools adapted to the task of information extraction from historical text, whereas tools developed for the modern language will generally not perform well when applied to historical text, due to differences and inconsistencies in spelling, vocabulary, morphology, and syntax as compared to standard modern language.

The field of natural language processing for historical text poses a challenging and interesting task for computational linguists to tackle, and in recent years, there has been a growing interest in this area. 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). However, computational linguists are not always familiar with the actual needs experienced by researchers in humanities. Likewise, historians and other researchers in humanities are often not aware of the possibilities offered by language technology tools and methods. Therefore, an important task within digital humanities is to bring the fields of computational linguistics and humanities research closer to each other, to encourage fruitful cooperation that would be beneficial for both fields.

In this paper, we present the outcome of a close collaboration between the fields of computational linguistics and history, resulting in a graphical user interface for information extraction from historical text. We present the underlying language technology tools and methods used in the system, as well as the graphical user interface provided to the historians. In addition to automatic evaluation measures of system performance, one of the main contributions of this paper is a thorough user evaluation of the extraction tool, performed by the historians, taking both quantitative and qualitative aspects into account.

The specific extraction task targeted in this study, is the extraction of verb phrases describing work from Early Modern Swedish text (approximately 1550–1800). This particular information need has arisen within the *Gender and Work* project at the Department of History at Uppsala University, where researchers are exploring what men and women did for a living in the Early Modern Swedish society. Within this project, phrases containing information about male and female working activities have so far been manually sought for in court records and other documents from the time period in question, and stored in a database referred to as the *Gender and Work database*. During this work, it has been noticed that working activities are most often expressed in the form of verb phrases, such as *to fish herring* or *to sell clothes* [**7**].

The main goal of the interface presented in this paper, henceforth referred to as *HistSearch*, is to facilitate the otherwise time-consuming manual search for evidence of working activities in historical input text, by automatically extracting phrases that are likely to describe work, and present these to the historians for further analysis. The automatic extraction workflow developed for this purpose is based on morphological and syntactic analysis using state-of-the-art taggers and parsers available for the standard modern language. To handle spelling variation in the historical input text, a spelling normalisation step is included as a core component in the pipeline, transforming the original spelling into a more modern spelling, before applying the tagger and the parser. From the linguistically annotated text, all the phrases analysed as verb phrases are extracted, based on the annotation labels.

In the final step, the extracted verb phrases are ranked, using statistical machine learning methods, so that those verb phrases that are most likely to describe working activities are displayed at the top of the results list. The ranked list of verb phrases is then displayed to the historians in a web-based interface, with possibilities to click on certain phrases to view them in context.

In Section 2, we present the Gender and Work project in more detail, and also discuss previous work on natural language processing for historical text. In Section 3, we describe further each step in the information extraction approach, whereas Section 4 is focused on the graphical user interface. Section 5 presents both an automatic evaluation of system performance, and a manual evaluation from the point of view of the historians. Finally, conclusions are drawn in Section 6.

2 Related Work

2.1 The Gender and Work Project

The Gender and Work Project is a combined research and digitization project with the aim of increasing knowledge about the ways in which men and women made a living in the past. The project has resulted in research publications [**26**], as well as an online database, where source fragments describing working activities have been collected, classified and made searchable (gaw.hist.uu.se).

Although fundamental to our understanding of both economic development and gender relations, knowledge of the gender division of work in Early Modern societies has been limited because of a lack of valid and systematic data [**15**,**10**,**24**]. Occupational designators are rare, especially for women, and many times vague or deceptive. Most work was unpaid and, hence, not noted in payrolls and accounts. Data on property – found, for example, in tax registers and probate inventories – say little about who actually did what work. More informative, and more valid, are descriptions of actual activities, performed by particular individuals at particular occasions. Many sources contain such information, but this information is scattered and often accidental. It appears at various rate of recurrence in court records, letters, diaries, etc. A witness in a court case may state, for instance, that she was reaping, when she saw what happened, or a diarist may have written down that 'today, I have spent most of my time spinning' [**7**].

A systematic and large-scale collection of such observations of everyday activities would provide a firm basis for conclusions about what men and women really did for a living in the past. Up until today, about 22,000 verb phrases describing working activities have been manually extracted, classified and stored in the Gender and Work database, combined with information on who performed the activity, when and in what context. While this dataset is relatively large compared to most historical research, it still represents only a small fragment of all the relevant sources from Sweden between 1550 and 1800. Collecting this kind of data is difficult and extremely time-consuming. Developing new methods allowing for faster extraction of relevant information from large amounts of text, in a systematic and reliable way, would be crucial for this line of research.

2.2 NLP for historical text

In the context of natural language processing (NLP) for historical text, one key issue that needs to be handled is the different and inconsistent spelling in historical text. Piotrowski [**19**] points out that historical text exhibits both diachronic and synchronic spelling variance, where diachronic spelling variance refers to the fact that languages change over time, leading to differences in spelling between different time periods. Synchronic spelling variance on the other hand, refers to inconsistencies in spelling between texts written in the same time period, or even within the same text written by the same author. These inconsistencies are mainly due to the lack of spelling conventions in the past, causing writers to spell words the way they sound, which is a subjective assessment, with dialectal and individual variance. In previous work on NLP for historical text, there are two main approaches to deal with the spelling issue. Either, the tools are adapted to the data, or the data are adapted to the tools.

In the first case, taggers and parsers are trained on linguistically annotated historical text, resulting in NLP tools adapted to the specific domain of historical text. One example of such an approach is the bootstrapping procedure implemented by Rögnvaldsson and Helgadóttir [**23**] for adapting the TnT tagger [**3**] to Old Icelandic.

For the reverse approach, that is to adapt the data to the tools, the key component is *spelling normalisation*. The main idea of spelling normalisation is to automatically convert the original, historical spelling to a more modern spelling, before applying taggers and parsers to the text. This way, existing NLP tools developed for the modern language may be used for analysing historical input data, avoiding the laborious work of developing NLP tools adapted to the task of analysing historical text. Several normalisation methods have been explored, often inspired by other subfields of computational linguistics, such as spell checking, speech technology, and machine translation.

One approach related to spell checking, is to perform edit distance comparisons between the historical word form and modern word forms, with the aim of substituting the historical word form by the modern word form that is closest in spelling to the original word form. One such approach was implemented by Bollmann [**2**] for normalising Early New High German (14th to 16th century) into Modern German. Similarly, Pettersson et al. [**17**] also tried an edit distance approach for normalising Early Modern Swedish into Modern Swedish.

Jurish [**11**] implemented an alternative method for spelling normalisation referred to as *conflation by phonetic form*, in which spelling normalisation is performed on the basis of phonetic similarity rather than orthographic similarity, the assumption being that phonetic properties are less resistant to diachronic change than orthography.

Spelling normalisation could also be viewed as a translation problem, where the historical language is to be translated into the modern language. Pettersson et al. [**18**] explored statistical machine translation (SMT) methods for spelling normalisation. In order to translate differences in spelling rather than the full translation of words and phrases, they performed translation at a character level, a method that has previously been used for translation between closely related languages.

3 System Overview

The HistSearch tool presented in this paper builds on the modular workflow for information extraction from historical text presented in Pettersson [**16**], as illustrated in Figure 1 and further described below. In the first step, the historical input text is tokenised, using standard tokenisation methods. Thereafter,

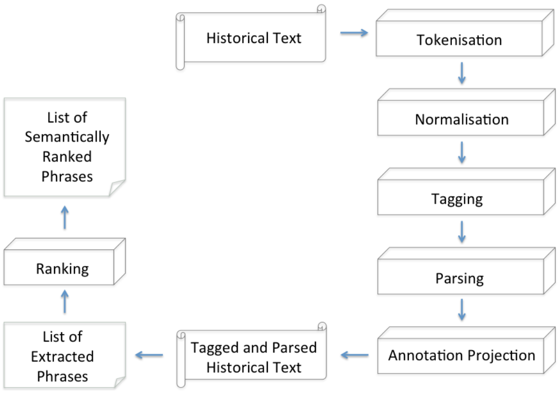

Historical Text
Tokenisation
Normalisation
Tagging
Parsing
Annotation Projection
Tagged and Parsed Historical Text
List of Extracted Phrases
Ranking
List of Semantically Ranked Phrases

Figure 1: HistSearch – System Overview

each token is normalised to a more modern spelling, to enable the application of modern NLP tools for the subsequent linguistic analysis step. The normalisation method used in the HistSearch tool builds on character-based statistical machine translation, as described further in Section 2.2 and in Pettersson et al. [**18**].

After normalisation, the input text is morphologically analysed using the *HunPos* tagger [**8**], with a pre-trained language model based on the *Stockholm-Umeå Corpus*, version 2.0 [**6**]. For syntactic analysis, the dependency parser *MaltParser* version 1.6 is used [**13**], with a pre-trained model based on the *Talbanken* section of the *Swedish Treebank* [**14**]. The annotation labels suggested by the tagger and the parser are then projected back from the normalised version of the text to the text in its original spelling, resulting in a tagged and parsed version of the historical text, with its original spelling preserved.

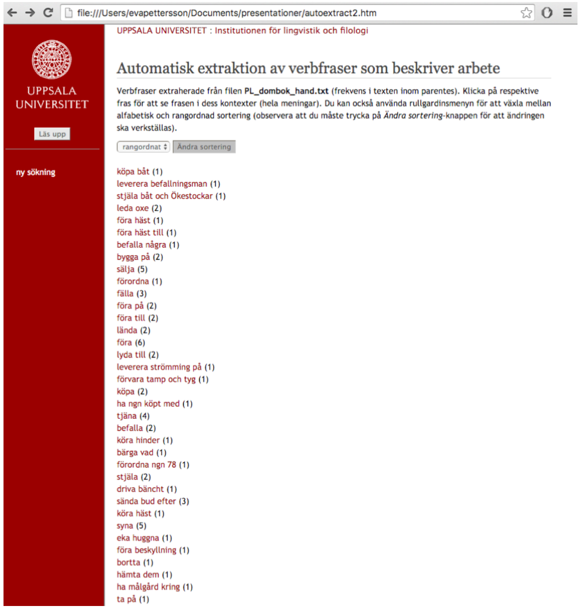
From the tagged and parsed version of the input text, verb phrases are retrieved by extracting all the words analysed as verbs by the tagger, together with all the phrases analysed as any of the following complements connected to the verb in question: subject (for passive verbs only), direct object, indirect object, predicative complement, prepositional complement, infinitive complement of object, verb particle, or reflexive.

Finally, the list of extracted verb phrases is ranked, so that those phrases that are most likely to describe working activities are presented at the top of the results list, whereas the phrases that are least likely to describe working activities are presented at the bottom of the results list. This ranking task is performed using bag-of-words classification in the Weka data mining software package, version 3.6.10 (Hall, et al., 2009), with a support vector machine (SVM) classifier with the sequential minimal optimisation (SMO) algorithm as defined by Platt [**21**]. As training data, we make use of a subset of the Gender and Work corpus, containing phrases stored in the database as describing work and linked to the source text in which they were found. By using the HistSearch pipeline for extracting all the verb phrases from these source texts, we get a set of positive instances in the form of the phrases stored in the database as describing work, and a set of negative instances in the form of phrases analysed as verb phrases by HistSearch but not stored in the database, thus assumed **not** to describe work. From these data, we train a classifier in Weka outputting the probability that a certain phrase describes work, based on the word forms contained in the phrase (more specifically based on the word forms analysed as verbs and nouns in the phrase, see Pettersson et al. [**16**] for further details).

4 Graphical User Interface

To enable testing and evaluation of the usefulness of HistSearch, a prototypical graphical user interface was developed. From this interface, the user is first prompted to upload a plain text file for the system to process. When clicking the *upload* button, the file is tokenised, normalised, tagged, and parsed in accordance with the system description presented in Section 3. In the next step, the verb phrases are extracted based on the annotation labels given by the tagger and the parser, and ranked so that those phrases that are likely to describe working activities are displayed at the top of the results list.

Figure 2 shows the resulting list of ranked verb phrases extracted from a snippet of *Per Larssons dombok*, a court records text from 1638 [**5**]. In the user interface, each element in the list is a normalised and lemmatised segment extracted from the input text and analysed as a verb phrase by HistSearch. Next to each phrase is a frequency number, indicating the number of times the specific phrase has been identified in the input text.


file:///Users/evapettersson/Documents/presentationer/autoextract2.htm
UPPSALA UNIVERSITET
Läs upp
ny sökning

UPPSALA UNIVERSITET : Institutionen för lingvistik och filologi

Automatisk extraktion av verbfraser som beskriver arbete

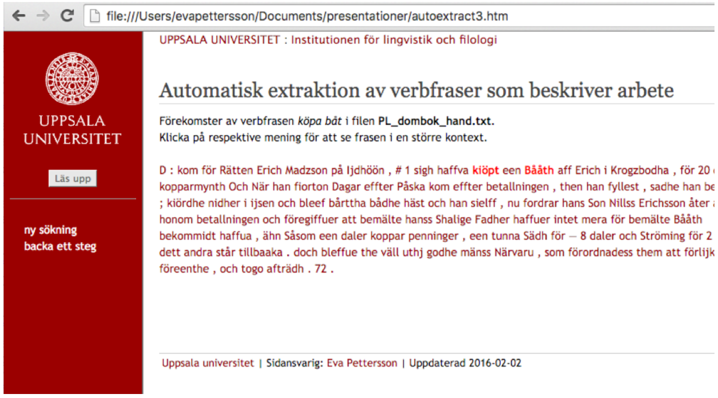
Verbfraser extraherade från filen **PL_dombok_hand.txt** (frekvens i texten inom parentes). Klicka på respektive fras för att se frasen i dess kontexter (hela meningar). Du kan också använda rullgardinsmenyn för att växla mellan alfabetisk och rangordnad sortering (observera att du måste trycka på *Ändra sortering*-knappen för att ändringen ska verkställas).

rangordnat ⬍ | Ändra sortering

köpa båt (1)
leverera befallningsman (1)
stjäla båt och Ökestockar (1)
leda oxe (2)
föra häst (1)
föra häst till (1)
befalla några (1)
bygga på (2)
sälja (5)
förordna (1)
fälla (3)
föra på (2)
föra till (2)
lända (2)
föra (6)
lyda till (2)
leverera strömming på (1)
förvara tamp och tyg (1)
köpa (2)
ha ngn köpt med (1)
tjäna (4)
befalla (2)
köra hinder (1)
bärga vad (1)
förordna ngn 78 (1)
stjäla (2)
driva bäncht (1)
sända bud efter (3)
köra häst (1)
syna (5)
eka huggna (1)
föra beskyllning (1)
bortta (1)
hämta dem (1)
ha målgård kring (1)
ta på (1)

Figure 2: Example of a ranked list of verb phrases extracted by HistSearch and displayed in the graphical user interface

The user also has the possibility to click on each phrase, to view it in a larger context, with the historical spelling preserved. This is exemplified in Figure 1, displaying the single occurrence of the phrase *köpa båt* (‘to buy a boat’), realized as *kiöpt Bååth* (‘bought Boat’), in the input text.


file:///Users/evapettersson/Documents/presentationer/autoextract3.htm
UPPSALA UNIVERSITET
Läs upp
ny sökning
backa ett steg

UPPSALA UNIVERSITET : Institutionen för lingvistik och filologi

Automatisk extraktion av verbfraser som beskriver arbete

Förekomster av verbfrasen *köpa båt* i filen **PL_dombok_hand.txt.**
Klicka på respektive mening för att se frasen i en större kontext.

D : kom för Rätten Erich Madzson på Ijdhöön , # 1 sigh haffva kiöpt een Bååth aff Erich i Krogzbodha , för 20
kopparmynth Och När han fiorton Dagar effter Påska kom effter betallningen , then han fyllest , sadhe han be
; kiördhe nidher i ijsen och bleef bårttha bådhe häst och han sielff , nu fordrar hans Son Nillss Erichsson åter
honom betallningen och föregiffuer att bemälte hanss Shalige Fadher haffuer intet mera för bemälte Bååth
bekommidt haffua , ähn Såsom een daler koppar penninger , een tunna Sädh för — 8 daler och Ströming för 2
dett andra står tillbaaka . doch bleffue the väll uthj godhe mänss Närvaru , som förordnadess them att förlijk
föreenthe , och togo afträdh . 72 .

Uppsala universitet | Sidansvarig: Eva Pettersson | Uppdaterad 2016-02-02

Figure 1: The verb phrase *köpa båt* ('to buy a boat'), realized as *kiöpt Bååth* (‘bought Boat’), displayed in a larger context in the graphical user interface

It should be noted that the user interface is only a simple prototype, which has not really been tested until now.

5 Evaluation

Evaluation of the HistSearch tool was performed both automatically and manually. In the automatic evaluation, *precision at k*, *R-precision*, and *average precision* were calculated, whereas the manual evaluation was performed by the historians, taking both quantitative and qualitative aspects into account. In the manual evaluation, three different methods for extraction of verb phrases describing work were tested and compared to each other, two of which used the HistSearch tool in order to extract the relevant verb phrases:

A. The researcher read the text from the beginning to the end, in the traditional way, and extracted manually relevant verb phrases as they turned up in the source.

B. The researcher uploaded the text in HistSearch and, using the ranked list, decided whether a suggested verb phrase was relevant or not by clicking on the phrase and looking at it in a larger context.

C. The researcher uploaded the text in HistSearch and read the text bit by bit by clicking on suggested verb phrases in the order of the ranked list. In contrast to Method B, the user was however not confined to the specific verb phrase suggested by the system, but extracted all relevant verb phrases found in the same text passage pointed to by the system.

In order to reduce the impact of differences between individual researchers, the source text was cut in half so that one researcher worked with method A for the first half of the text and method B for the second half, while another researcher did vice versa. Method C on the other hand, was performed by one researcher only (a third researcher).

5.1 Data

As input text for the evaluation an extract from the court records of the magistrate of Linköping, written in 1709, was used. The extract covers three court sessions from February to March 1709, and consists of 5,792 words, corresponding to 32,547 characters. The text was transcribed manually from the original document (but the transcription was not checked). When uploaded in HistSearch, the system generated 881 verb phrases in total from the input text. To decide which of these phrases that describe working activities is not always a trivial task. It is a matter of interpretation and, sometimes, discussion. A pragmatic criterion was used: If at least two of the three researchers considered a suggested verb phrase to be relevant, it was counted as a 'hit'. By using this criterion, a gold standard was created. In all, the gold standard consisted of 70 verb phrases. This amounts to 7.9% of all listed verb phrases, or one instance per 83 words.

5.2 Time-use

One goal in the development of the HistSearch tool has been to make the verb phrase extraction task faster, by presenting the automatically analysed phrases to the historian in a ranked list with the phrases that are most likely to describe work presented at the top of the results list, meaning that the researcher would not need to read the whole source document to find the phrases of interest. To see if this goal has been achieved, the time used for extracting relevant phrases from the evaluation text, using the three different methods presented above, was calculated.

To read the source text and manually extract relevant verb phrases (Method A) took 57 minutes (in total, for both researchers combined). By contrast, when uploading the file in HistSearch, the automated steps of tokenisation, normalisation, tagging, parsing, verb phrase extraction and ranking were completed in 45 seconds, that is only a very small fraction of the time needed by a human reader.

Deciding whether a suggested verb phrase was relevant or not took, on average, 11 seconds. With this speed, it would take 18 minutes to go through the top-100 ranked verb phrases. Considering that some of the listed verb phrases were very easily dismissed, as they already at first glance clearly did not describe a working activity, a top-100 list consisting of many relevant verb phrases would in reality take longer to go through, though probably no more than 30 minutes.

Going through the whole list, and checking each of the 881 verb phrases (Method B) was done in 165 minutes, or 2 hours and 45 minutes. As stated, only a fraction of the total number of listed verb phrases was however deemed relevant. This points to the importance of an effective and reliable ranking. With a good ranking approach, the bottom part of the results list could be skipped or only skimmed through by the researcher.

Method C, i.e. reading the text piecemeal, in the order of the ranked list by clicking on suggested verb phrases, was more or less equal in time to Method B (2 hours and 46 minutes). One reason why this method was much more time-consuming than reading the source text from the beginning to the end, was that the list of verb phrases directed the reader to the same text passages several times. Another reason could be that the historians are already very familiar with Method A, which has been performed many times before, whereas using the interface is a new experience which may take some time to get used to. Yet another reason is that reading an 18th century court record in a traditional, linear, way helps interpretation. We return to this in section 6.

5.3 Ranking Results

The ranking part of the information extraction process was automatically evaluated in terms of precision at k, R-precision, and average precision. *Precision at k* is defined as the precision at certain positions in the ranked list [**12**]. For example, precision at 10 measures the proportion of relevant instances displayed among the top-10 instances in the list. *R-precision* (also referred to as the *break-even point*) is a special case of precision at k, referring to the precision at the position in the list where the number of extracted instances is equal to the number of relevant instances, at which point precision and recall are the same [**4**]. *Average precision* is calculated as the average of the precision for all positions in the list where relevant instances are found, until all relevant instances have been retrieved [**25**]. The automatic evaluation results for the input text presented in Section 5.1 are given in Table 1.

	p@10	p@50	p@100	R-precision	AVP
Random	0.00	0.10	0.07	0.09	0.08
Ranked	0.40	0.34	0.24	0.30	0.22

Table 1: Automatic evaluation results for ranking verb phrases describing work in the HistSearch tool. p@10 = precision at 10. p@50 = precision at 50. p@100 = precision at 100. AVP = Average Precision.

When compared to the randomised list, where the extracted verb phrases are simply ordered alphabetically, the ranked list is helpful. In the alphabetically ordered list, no relevant verb phrases were found among the top-10 instances in the list, as compared to 4 relevant phrases in the ranked list. Likewise, only 7 out of the top-100 instances in the alphabetically ordered list were relevant instances, as compared to 24 phrases in the ranked list.

The results are however still rather low, especially considering previous evaluation results presented in Pettersson [**16**], where 8 out of the top-10 instances where relevant instances, using the same pipeline for extraction of verb phrases describing work from Early Modern Swedish text. Those experiments were however performed on a different kind of data set, not containing full source texts but only source text snippets with a high degree of relevant verb phrases, both in the training data and in the test data. For comparison, 27% of the verb phrases occurring in the previous training and evaluation sets were relevant verb phrases, as compared to 7.9% in the current evaluation text. Since machine learning methods are sensitive to the proportion of positive and negative examples in the training data, it is therefore likely that the ranking approach when applied to real-life texts could be approved by training on full source texts rather than non-representative source text snippets.

In the current setting, if the top-100 instances had been consulted, only one third of the relevant verb phrases in the source text would have been detected. The number of 'false negatives', that is verb phrases that are ranked low even though describing working activities, is thus high in this sample. In order to find two thirds, the researcher would have had to look at 426 verb phrases in the list, which would have taken 80 minutes (assuming an average time of 11 seconds per verb phrase), that is a longer time than was spent on reading the text in the traditional way with manual extraction of verb phrases. Some 'false positives", that is verb phrases that were highly ranked even though not describing working activities, can be explained by ambiguity. For example, while the Swedish word *bära* (just as its English equivalent 'carry') describe work in some, or perhaps most, contexts (as in 'carry a sack of corn'); in others, it has nothing to do with work, as in *bära sig åth* ('behave').

5.4 Redundance and Absence

Using the HistSearch tool for extracting verb phrases (Method B and C) generated more verb phrases than the manual extraction (Method A). This was true for all users. One reason is that the system returned some of the verb phrases twice, or more, in somewhat different forms, and the user did not always observe this. For example, the phrase *har pantsatt sin gård* ('has pledged his house') was listed under both 'pledge house' (at the top) and 'have' (further down).

Sometimes, the court record described an activity in two different ways. In one court case, two women were described as having inspected (*besichtigat*) a dead child. Later on, they were said to have been to see the child (*warit och sedt barnet*). To the reader who has read the case from the beginning, it is obvious that these two expressions describe one and the same activity. It may not be equally obvious for the interface user, who finds herself thrown into the middle of a text passage on two different occasions.

Another, more positive explanation to why using HistSearch generated more verb phrases would be that an interface user is more consistent in extracting instances of the same kind, as similar phrases are grouped together in the ranked list, regardless of their position in the original text. For example, several instances of 'issue bill of sale' were treated more consistently by the interface user.

Using HistSearch may cause some redundance in the list of verb phrases, then. This is not a serious problem, especially if it is the effect of a more consistent way of collecting data. In the end, all verb phrases need to be checked manually before being registered in the Gender and Work database. To sort out a few superfluous verb phrases – false positives – would not cause considerable extra work. Absence of relevant verb phrases – false negatives – is a more serious problem. There were a few verb phrases identified by the manual reader, but missed by the system: one instance of 'buy tobacco' (*Kiöpt toback*) and one of 'sell bricks' (*bortsåldt teglet*). Given the training data, which includes many instances of buying and selling, it could have been expected that the system would extract these phrases and rank them high. In this particular case however, *kiöpt* ('have bought') was analysed as a past participle by the tagger rather than a past tense verb, which is the reason why it was not extracted as belonging to a verb phrase. In the case of *bortsåldt* ('have sold'), this is a verb particle construction, comprised of the particle *bort* ('away') merged with the verbform *såldt* ('sold'). In present-day Swedish the particle would not be part of the verb form, but instead not used at all, or possibly succeeding the verb as a stand-alone particle: *sålt bort* ('sold away'). Since the merged word form is thus unknown to the tagger, even after correct spelling normalisation, it has been analysed as a proper noun instead of a verb. The high variability of historical data will always lead to data sparseness problems of this kind, meaning that we could never expect the NLP tools to perform flawlessly. The aim is however to be able to extract most of the relevant instances, and to be able to rank them with high confidence. It could also be noted that since the HistSearch tool offers the possibility to view all verbs in a larger context, the researcher who used HistSearch, but was not limited to extracting verb phrases identified by the system (Method C), did not miss these instances.

6 Conclusion

In this paper, we have presented a first version of the HistSearch tool for automatic extraction of verb phrases describing work from Early Modern Swedish text. This tool is the outcome of intersecting research interests of two disciplines. For the computational linguists, the aim is to explore how language technology tools could be used for analysing historical text. For the historian, it is a question of developing more efficient, and more reliable, methods for data collection.

Both from the automatic and the manual evaluation results presented in this paper, it is clear that there are still issues to be solved. HistSearch is no ready solution, but its modular workflow provides a platform enabling further development of specific parts of the pipeline, to make the tool more accurate and useful.

The most important, and most successful, step taken so far is the normalisation of spelling, which allows for automatic linguistic analysis of historical text, such as lemmatisation and tagging, using existing NLP tools developed for the modern language. In the area of parsing and ranking, more work needs to be done. While the system performs well when it comes to identifying a verb, it more seldom points out its relevant complements. Thus, parsing performance needs to be improved. Currently, only spelling is considered in the normalisation step. For better parsing results, syntactic differences (and inconsistencies) in historical text as compared to modern text, such as word order differences, would also need to be handled. Future work further includes better ranking of the extracted verb phrases. As mentioned in Section 5.3, the current ranking system was trained on rather unrepresentative text, containing a larger proportion of verb phrases describing work than an average input text would do. It is thus assumed that the ranking step could be improved by retraining on more representative texts. Another way of improving the ranking step would be to include interactive retraining of the ranking system. This could be done through the user interface, by allowing for the user to judge the suggested phrases as relevant or not, whereby the system would add these judgements to the sets of positive and negative examples for ranking. In this way, the ranking system would learn from the user input, and gradually improve its ranking strategies.

In developing a new user interface, it would be important to make more use of the domain knowledge of the historians, who are familiar with the structure and content of historical texts. The historian who has studied many sources of a similar kind has an expertise knowledge which is not trivial to express in the form of algorithms that a computer could process. Working with the graphical user interface makes it clear that approaching a text from a list of presumably relevant text passages is something very different from reading a text in a traditional manner. Court records are not only a collection of words or word strings, but include a series of narratives that need to be reconstructed by the researcher. This is not always easily done from snippets. It matters where in the text a certain phrase occurs. In the beginning of a court session, for example, records often include the proclamation of royal or local ordinances. These are seldom of interest to the historian who is interested in working activities, which affects the reading. By contrast, the verb phrase ranked as number one in the list generated by the system is a general statement from such a context. Thus, another way of improving the user interface would be to observe the order of the text and to facilitate overview of the entire text.

It should be noted that the evaluation presented in this study was performed on a rather small test text. For longer texts, the manual extraction process would be much more time-consuming, and provided that the HistSearch tool is further improved in terms of parsing performance and ranking strategies, the automatic extraction process could be a very useful method for extracting large amounts of information from historical text in a shorter period of time.

References

1 Black, A.W. and Taylor, P. *The festival speech synthesis system: system documentation*. Human Communciation Research Centre, University of Edinburgh. 1997.

2 Bollmann, M. POS Tagging for historical texts with sparse training data. in *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse* (2013), 11-18.

3 Brants, T. TnT – a statistical part-of-speech tagger. in *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)* (2000), 224-231.

4 Craswell, N. R-precision. in *Encyclopedia of database systems*. Springer US, Boston, 2009.

5 Edling, N. *Uppländska domböcker 4: lagläsaren Per Larssons dombok 1638*. Almqvist & Wiksells, Uppsala, 1937.

6 Ejerhed, E. and Källgren, G. *Stockholm Umeå Corpus. Version 1.0*. Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. 1997.

7 Fiebranz, R., Lindberg, E., Lindström, J., and Ågren, M. Making verbs count: the research project "Gender and Work" and its methodology. *Scandinavian Economic History Review*, 2, 59 (2011), 271-291.

8 Halácsy, P., Kornai, A., and Oravecz, C. HunPos - an open source trigram tagger. in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (2007), 209-212.

9 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. The WEKA data mining software: an update. *SIGKDD Explorations* (2009), 10-18.

10 Humphries, J. and Sarasúa, C.. Off the record: reconstructing women's labor force participation in the European past. *Feminist economics* (2012), 39-67.

11 Jurish, B. Finding canonical forms for historical German text. in *Text resources and lexical knowledge: selected papers from the 9th Conference on Natural Language Processing (KONVENS)*. Mouton de Gruyter, Berlin, 2008.

12 Manning, C., Raghavan, P., and Schütze, H.. *Introduction to information retrieval*. Cambridge University Press, Cambridge, 2008.

13 Nivre, J., Hall, J., and Nilsson, J.. MaltParser: A data-driven parser-generator for dependency parsing. in *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)* (2006), 2216-2219.

14 Nivre, J., Nilsson, J., and Hall, J. Talbanken05: a Swedish treebank with phrase structure and dependency annotation. in *Proceedings of the 5th international conference on Language Resources and Evaluation (LREC)* (2006), 24-26.

15 Ogilvie, S.. *A bitter living: women, markets, and social capital in early modern Germany*. Oxford University Press, Oxford, 2003.

16 Pettersson, Eva. *Spelling normalisation and linguistic analysis of historical text for information extraction*. Uppsala University, Dissertation, 2016.

17 Pettersson, E., Megyesi, B., and Nivre, J. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)* (2013), 163-179.

18 Pettersson, E., Megyesi, B., and Tiedemann, J.. An SMT approach to automatic annotation of historical text. in *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA* (2013), Linköping Electronic Conference Proceedings, 54-69.

19 Piotrowski, M.. *Natural language processing for historical texts*. Morgan & Claypool Publishers, 2012.

21 Platt, J. C. *Sequential minimal optimization: a fast algorithm for training support vector machines*. 1998.

22 Rayson, P., Archer, D., and Smith, N. 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* (2005).

23 Rögnvaldsson, E. and Helgadóttir, S. Morphosyntactic tagging of old Icelandic texts and its use in studying syntactic variation and change. in *Theory and applications of natural language processing*. Springer, Berlin Heidelberg, 2011.

24 Whittle, J. Enterprising widows and active wives: women's unpaid work in the household. *The History of the Family* (2014), 283-300.

25 Zhang, E. and Zhang, Y. Average precision. in *Encyclopedia of database systems*. Springer US, Boston, 2009.

26 Ågren, Maria. *Making a living, making a difference: gender and work in early modern society*. Oxford University Press, New York, In print.