**Automatic verb extraction from historical Swedish texts**

Eva Pettersson  
eva.pettersson@convertus.se  
May 12, 2010

**Contents**

1 Introduction .................................................. 2

2 Creating a gold standard for evaluation .................. 2

3 Normalisation and morphological analysis ............... 3
   3.1 SALDO ....................................................... 4
   3.2 Morphological analysis using SALDO ................. 4
      3.2.1 Initial normalisation .............................. 5
      3.2.2 Stepwise normalisation ........................... 5
      3.2.3 Morphological analysis of modern Swedish text .. 6

4 Combining morphological analysis and tagging .......... 7
   4.1 The HunPOS tagger ...................................... 7
   4.2 Combining SALDO and HunPOS ......................... 7

5 Analysing OCR-scanned text as compared to manually entered text .................. 8

6 Extracting relevant verb forms .......................... 10

7 A Graphical User Interface for automatic verb extraction 12
   7.1 Uploading files .......................................... 13
   7.2 Generating a list of verbs ............................ 13
   7.3 Verb concordances ...................................... 13
   7.4 A larger context ........................................ 14

8 Conclusions .................................................. 14

9 Further development ......................................... 15

A Normalisation rules for the morphological analyser (in the order of application) .... 18

B Normalisation rules for the tagger ......................... 19
1 Introduction

The aim of the project "Gender and Work" is to study what men and women did for a living in Sweden during the period 1550-1800. This information is revealed in a large number of historical texts, and a time-consuming part of the project is to manually go through historical documents to pick out the relevant text sequences and store these in the Gender and Work database (GaW), a large-scale relational database developed within the project.

In the project, it has been noticed that the information on what people did for a living is often formulated as a transitive verb with a direct object. Since the task of manually finding relevant verb constructions in a text is very time-consuming, it is interesting to explore the idea of facilitating this task, using existing natural language processing techniques. However, such tools are developed for handling modern Swedish written language, and it is yet to be explored how these tools perform on historical texts.[1]

This report describes an attempt to facilitate the extraction process, using morphological analysis tools. The aim is to explore whether it is possible to automatically extract the verbs in historical Swedish texts using modern natural language processing tools, and further, what tools are appropriate for this specific task, and what level of accuracy that may be achieved. An additional aim is to examine if and how the results are affected when analysing OCR-scanned text as compared to manually entered text. Another question addressed is how to integrate the verb extraction techniques in a user interface. Furthermore, in the Gender and Work context, only a subset of all verbs are relevant to the historians, wherefore it is also interesting to study the results of identifying these verbs only.

Section 2 describes the creation of a gold standard for evaluation. In section 3 morphological analysis results for a historical text is presented, and how normalisation of the input affects the quality in terms of precision and recall. In section 4, the results of combining morphological analysis and tagging are presented, while section 5 focuses on analysing OCR-scanned text as compared to manually entered text. Section 6 discusses how precise the extraction techniques are in finding excerpts of interest, and section 7 describes the development of a prototypical graphical user interface, integrating the automatic verb extraction techniques developed. Finally, some conclusions are drawn in section 8, and further development issues are discussed in section 9.

2 Creating a gold standard for evaluation

The text used as a basis for developing the automatic verb extraction tool described in this report, is a subset of Per Larssons dombok, a selection of court records from 1638[4]. This text consists of 11 439 tokens in total, and is available both in an OCR-scanned version, and a manually entered version. A gold standard for evaluation was created, by manually annotating all the verbs in the manually entered version of the text. This verb annotation procedure was performed using an annotation scheme describing four different verb types (main verbs, temporal auxiliary verbs, modal auxiliary verbs and copula verbs) and four different verb forms (finite, infinitive, imperative and supine), yielding a total of 16 possible verb tags. Participle forms were not
defined as verbs, as participles are normally used in an adjectival function rather than a verbal one. The complete annotation scheme is listed in table 1.

| verb tag | description                  |
|----------|------------------------------|
| VBFM     | main verb, finite            |
| VBIM     | main verb, infinitive        |
| VBMM     | main verb, imperative        |
| VBSM     | main verb, supine            |
| VBFT     | temporal auxiliary verb, finite |
| VBIT     | temporal auxiliary verb, infinitive |
| VBMT     | temporal auxiliary verb, imperative |
| VBST     | temporal auxiliary verb, supine  |
| VBFO     | modal auxiliary verb, finite  |
| VBO      | modal auxiliary verb, infinitive |
| VBMO     | modal auxiliary verb, imperative |
| VBSO     | modal auxiliary verb, supine  |
| VBFC     | copula verb, finite          |
| VBIC     | copula verb, infinitive      |
| VBMC     | copula verb, imperative      |
| VBSC     | copula verb, supine          |

Table 1: Verb annotation scheme for creating the gold standard

For the verb annotation to be as accurate as possible, the text was annotated independently by two persons (Eva Pettersson and Martin Palm, Department of Linguistics and Philology, Uppsala University). The parallel annotation was performed in three steps, and the results were analysed and compared between each annotation round.

In the first round, the initial 984 tokens were annotated. A comparison between the two independently performed annotations revealed some differences, mainly concerning what words should be regarded as participles rather than verbs.

Having reached consensus for the first part of the text, a new parallel annotation was performed of the succeeding 954 tokens. This resulted in a far better consensus, even though there was still some disagreement, mainly concerning participles, and the division into main verbs versus auxiliary verbs.

In the third round, all the remaining tokens were annotated. The results of the annotation process are presented in table 2.

The resulting gold standard describes a text comprising 2 093 verbs in total, of which 1 357 are main verbs.

3 Normalisation and morphological analysis

This section presents the morphological analysis tool SALDO 1.0, and the normalisation efforts needed to use SALDO for analysis of historical texts.
Table 2: The manual verb annotation process

| Words marked as verb by person 1 | Round 1 | Round 2 | Round 3 |
|----------------------------------|---------|---------|---------|
| Words marked as verb by person 2 | 177     | 163     | 2088    |
| Words marked as verb by both persons | 156     | 154     | 2046    |
| Words marked with the same verb tag by both persons | 92      | 144     | 1893    |

3.1 SALDO

SALDO is a lexical resource developed for modern Swedish written language. It is based on Svenskt AssociationsLexikon (SAL), a semantic dictionary compiled in 1992 by Lennart Löngren (Centrum för datortjiisk and Slaviska institutionen, Uppsala University). In 2003, Lars Borin and Lennart Löngren (University of Gothenburg), initiated the work of making SAL digitally available, which eventually led to the development of SALDO.

The SALDO dictionary comprises 72,396 entries (07-05-2008) defined at lexeme level (68,355 lemma entries). The dictionary is based on SAL, combined with inflectional information as defined in Nationalencyklopedin, Svenska Akademiens Ordlista and Svenska Akademiens Grammatik. The following parts of speech are represented in the dictionary: nouns, proper nouns, adjectives, pronouns, numerals, verbs, adverbs, prepositions, conjunctions, subjunctions, the infinitive marker and interjections.

Apart from single word entries, the SALDO dictionary also contains approximately 2,000 multi-word units, including 1,100 verbs (mainly particle verbs). The SALDO package provides functions for morphological analysis, parsing and part-of-speech tagging. In the following, we will focus solely on the morphological analyser.

3.2 Morphological analysis using SALDO

Since the SALDO dictionary is adapted to modern Swedish written language, running a text with the old Swedish spelling preserved means that the analyser fails to assign a morphological analysis in many cases. When running the SALDO analyser on the test text, less than 50% of the words were analysed.

To increase the number of words found in the SALDO dictionary, I formulated some normalisation rules, transforming the text into a more modern spelling, before running the document through the morphological analyser.

The initial 984 tokens of the text were used as development data, i.e. words used as inspiration when formulating the normalisation rules, whereas the rest of the text was used solely for evaluation. Since the morphological analyser is context-insensitive, and the project aims at finding only the verbs, the main focus in formulating the rules has been on handling verb forms.
3.2.1 Initial normalisation

As a first step, 25 normalisation rules were formulated, and tested on three text sections in *Per Larssons dombok*. Each text was run through the morphological analyser both in its original form and after having applied the normalisation rules. The results show an increase in both precision and recall when the rules have been applied, as illustrated in table 3.¹

|                  | Tokens analysed | Verbs analysed as verbs | Verbs not analysed as verbs | Non-verbs analysed as verbs | Prec. | Recall |
|------------------|-----------------|-------------------------|-----------------------------|-----------------------------|-------|--------|
| unnorm. dev data | 762             | 116                     | 208                         | 132                         | 0.47  | 0.36   |
| normalised dev data | 1141            | 265                     | 59                          | 198                         | 0.57  | 0.82   |
| unnorm. test data | 737             | 104                     | 209                         | 118                         | 0.47  | 0.33   |
| normalised test data | 1105            | 255                     | 58                          | 182                         | 0.58  | 0.81   |
| unnorm. test data | 4324            | 601                     | 1168                        | 697                         | 0.46  | 0.34   |
| normalised test data | 6220            | 1315                    | 424                         | 1018                        | 0.56  | 0.76   |

Table 3: Morphological analysis results using SALDO 1.0, before and after applying normalisation rules

The first text section (*development data*) is the part of the text used as inspiration when formulating the normalisation rules. The second text section (*test data (1681 tokens)*), contains an equal amount of unseen text, whereas the third text section (*test data (9444 tokens)*), comprises all the unseen text in *Per Larssons dombok*, including the material in the second text section.

3.2.2 Stepwise normalisation

The initial set of normalisation rules was limited to rules that are rather safe to apply, in the sense that spelling is seldom transformed when it should not be transformed. One could therefore assume that recall could be increased, if more normalisation rules were defined. To increase the number of normalisation rules, without losing confidence in the rules, a stepwise normalisation procedure was introduced. In this process, the 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 normalisation step.

¹The total number of tokens is somewhat different from the number of tokens given in section 2, due to the fact that the word count function included in SALDO has a slightly different definition of words, compared to the bash command `wc` that was used in section 2.
Using this stepwise method, a total of 39 normalisation steps were defined, increasing recall from 76% to 86% as compared to the initial normalisation attempt described in section 3.2.1. Table 4 illustrates the results of the stepwise normalisation procedure, when run on the 9444 tokens test data.

|                  | Tokens analysed | Verbs analysed as verbs | Verbs not analysed as verbs | Non-verbs analysed as verbs | Prec. | Recall |
|------------------|-----------------|-------------------------|-----------------------------|-----------------------------|-------|--------|
| raw data         | 4324            | 601                     | 1168                        | 697                         | 0.46  | 0.34   |
| 5 rules          | 4660            | 806                     | 963                         | 742                         | 0.52  | 0.46   |
| 10 rules         | 6299            | 1274                    | 495                         | 1026                        | 0.55  | 0.72   |
| 15 rules         | 6450            | 1301                    | 468                         | 1059                        | 0.55  | 0.74   |
| 20 rules         | 6619            | 1406                    | 363                         | 1084                        | 0.56  | 0.79   |
| 25 rules         | 6948            | 1458                    | 311                         | 1145                        | 0.56  | 0.82   |
| 30 rules         | 6982            | 1479                    | 290                         | 1149                        | 0.56  | 0.84   |
| 35 rules         | 7120            | 1519                    | 250                         | 1171                        | 0.56  | 0.86   |
| 39 rules         | 7202            | 1530                    | 239                         | 1188                        | 0.56  | 0.86   |

Table 4: SALDO morphological analysis results, using a stepwise application of normalisation rules

The normalisation rules defined are inspired by (but not limited to) some of the changes in the reformed Swedish spelling introduced in 1906, where for example fv and hv should be rewritten as v (fördärfvat → fördärfvat), doubled vowels are transformed into a single vowel (söoka → söka) and dt is rewritten as t (varidt → varit) etc.

As a complement to the rules based on the spelling reform, a number of empirically designed rules were formulated, based on the development data described in section 3.2. The empirical rules include the rewriting of verbal endings (e.g. begårade → begårde and utviste → utvisade), transforming double consonants into a single consonant (vetta → veta, prööss → prööss) and vice versa (upsteg → uppsteg, visste → visste) etc. The total set of normalisation rules in the stepwise normalisation procedure is listed in appendix A.2

3.2.3 Morphological analysis of modern Swedish text

SALDO is developed for morphological analysis of modern Swedish text. To compare the results of running SALDO on historical documents to running SALDO on modern texts, SALDO was run on the Stockholm-Umeå Corpus (SUC), a modern 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.[5]

Running SALDO on SUC results in a very high level of recall, as illustrated in table 5.

2The normalisation rules in appendix A amounts to a few more rules than 39, as some of the rules describe the same change in spelling and are thus run in a single step.
4 Combining morphological analysis and tagging

The SALDO morphological analyser renders all analyses for a word form given in the dictionary, regardless of context. Due to ambiguous words, a high level of recall may thus be achieved, at the expense of low precision figures. 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.

To improve precision without losing too much in recall, one solution is to combine the morphological analysis with a tagger, to disambiguate the ambiguous words.

4.1 The HunPOS tagger

In parallel to my work on the morphological analyser, Martin Palm (Department of Linguistics and Philology, Uppsala University) has developed a set of normalisation rules for the HunPOS tagger [6] (a free and open source implementation of the TnT tagger [3]). The set of normalisation rules developed for the tagger includes a subset of the normalisation rules formulated for the morphological analyser, combined with further normalisation rules specifically adjusted to the tagger.

There are mainly two reasons why the same set of normalisation rules may not be used both for the morphological analysis and for the tagger. Firstly, 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, to acquire appropriate analyses for the verbs. Secondly, the tagger 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 a stepwise manner, as was proved to be successful for the morphological analyser, is not an option. Appendix B lists the set of normalisation rules used for the tagger.

4.2 Combining SALDO and HunPOS

I tried a combination of SALDO and HunPOS (trained on the SUC corpus [5]) for automatic verb extraction, comparing the results of four different methods for verb extraction:

1. a token is considered to be a verb if there is a morphological verb analysis
2. a token is considered to be a verb if it has been tagged as a verb

| Tokens | Tokens analysed | Verbs analysed as verbs | Verbs not analysed as verbs | Non-verbs analysed as verbs | Prec. | Recall |
|--------|----------------|------------------------|----------------------------|----------------------------|-------|--------|
| 1 022 775 | 93 807 | 173 521 | 1 353 | 155 859 | 0,53 | 0,99 |

Table 5: SALDO morphological analysis results for modern Swedish text (the SUC corpus)
3. A token is considered to be a verb if there is a morphological verb analysis or it has been tagged as a verb.

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

These four methods were evaluated on the test data from *Per Larssons dombok* (9,444 tokens). Table 6 presents the results of combining SALDO and HunPOS, running the tools on the test text in its original, unnormalised form, whereas Table 7 presents the results of combining SALDO and HunPOS, after normalisation of the text.

|                          | Verbs analysed as verbs | Verbs not analysed as verbs | Non-verbs analysed as verbs | Prec. | Recall |
|--------------------------|-------------------------|----------------------------|-----------------------------|-------|--------|
| Morph. analyser          | 601                     | 1168                       | 697                         | 0.46  | 0.34   |
| Tagger                   | 920                     | 849                        | 331                         | 0.74  | 0.52   |
| Morph. analyser or tagger| 1003                    | 766                        | 965                         | 0.51  | 0.57   |
| Morph. analyser and tagger| 518                    | 1251                       | 63                          | 0.89  | 0.29   |

Table 6: Recall and precision for unnormalised, manually entered text, combining morphological analysis and tagging

|                          | Verbs analysed as verbs | Verbs not analysed as verbs | Non-verbs analysed as verbs | Prec. | Recall |
|--------------------------|-------------------------|----------------------------|-----------------------------|-------|--------|
| Morph. analyser          | 1530                    | 239                        | 1188                        | 0.56  | 0.86   |
| Tagger                   | 1537                    | 232                        | 261                         | 0.85  | 0.87   |
| Morph. analyser or tagger| 1666                    | 103                        | 1318                        | 0.56  | 0.94   |
| Morph. analyser and tagger| 1401                   | 368                        | 131                         | 0.91  | 0.79   |

Table 7: Recall and precision for normalised, manually entered text, combining morphological analysis and tagging

5 **Analysing OCR-scanned text as compared to manually entered text**

So far, the verb extraction techniques have been evaluated on the manually entered version of the test text only. Since it is a time-consuming task to
manually digitalise historical texts, it would however be a great advantage to use OCR-scanning techniques in the digitalisation process. One could assume though, that OCR-scanned text contains more errors than manually entered text, due to erroneous character analyses. To explore how this potential source of error affects the identification of verbs, the verb extraction tool was run on the OCR-scanned version of the test text as well. The results show that any errors introduced by using OCR-techniques to digitalise a text did not have a noticeable effect on the retrieval accuracy, as illustrated in tables 6, 7, 8 and 9.

|                  | Verbs analysed as verbs | Verbs not analysed as verbs | Non-verbs analysed as verbs | Prec. | Recall |
|------------------|-------------------------|-----------------------------|-----------------------------|-------|--------|
| Morph. analyser  | 603                     | 1166                        | 695                         | 0.46  | 0.34   |
| Tagger           | 917                     | 852                         | 331                         | 0.73  | 0.52   |
| Morph. analyser  | 1003                    | 766                         | 962                         | 0.51  | 0.57   |
| or tagger        |                         |                             |                             |       |        |
| Morph. analyser  | 517                     | 1252                        | 64                          | 0.89  | 0.29   |
| and tagger       |                         |                             |                             |       |        |

Table 8: Recall and precision for unnormalised, OCR-scanned text, combining morphological analysis and tagging

|                  | Verbs analysed as verbs | Verbs not analysed as verbs | Non-verbs analysed as verbs | Prec. | Recall |
|------------------|-------------------------|-----------------------------|-----------------------------|-------|--------|
| Morph. analyser  | 1527                    | 242                         | 1181                        | 0.56  | 0.86   |
| Tagger           | 1528                    | 241                         | 264                         | 0.85  | 0.86   |
| Morph. analyser  | 1662                    | 107                         | 1315                        | 0.56  | 0.94   |
| or tagger        |                         |                             |                             |       |        |
| Morph. analyser  | 1393                    | 376                         | 130                         | 0.91  | 0.79   |
| and tagger       |                         |                             |                             |       |        |

Table 9: Recall and precision for normalised, OCR-scanned text, combining morphological analysis and tagging

Studying the results of the tagger on the manually entered version of the test text and on the OCR-scanned version, one could note that there are five unique verb forms in the text that are annotated differently in the two versions, where three verb forms are assigned a verb tag only in the manual version, and two verb forms are assigned a verb tag only in the OCR-scanned version. Furthermore, there are six unique non-verbs that are annotated as verbs only in the manually entered version of the text, as compared to three unique non-verbs annotated as verbs only in the OCR-scanned version. Table 10 summarises the
differences in verb annotation between the manually entered version and the OCR-scanned version of the text.

| OCR analysis                     | Manually ent. analysis | Gold standard |
|----------------------------------|------------------------|---------------|
| massa                            | māssa                  | non-verb      |
| noun indef singular              |                        |               |
| •                                |                        |               |
| verb infinitive                  |                        |               |
| paenam                           | paenam                 | non-verb      |
| unknown word                     |                        |               |
| samptlighe                       | samplighe              | non-verb      |
| noun + unknown word              |                        |               |
| Jāmvāl                           | Jāmvāl                 | non-verb      |
| interjection                     |                        |               |
| hanness                          | hannes                 | non-verb      |
| proper noun                      |                        |               |
| tillhörer                        | tillhöver              | verb          |
| verb present + pronoun           |                        |               |
| ordinarias                       | ordinariae             | verb          |
| verb present tense               |                        |               |
| hannes                           | hannes                 | non-verb      |
| verb past tense                  |                        |               |
| förre                            | förre                  | non-verb      |
| adjective positive masculine     |                        |               |
| bekant                           | bekant                 | verb          |
| adjective positive               |                        |               |
| stellless                        | stellles               | verb          |
| noun indefinite singular         |                        |               |
| handha                           | handha                 | verb          |
| verb infinitive                  |                        |               |
| Prof va                          | Pröfeva                | verb          |
| proper noun + interjection       |                        |               |

Table 10: Differences in verb annotation between OCR-scanned and manually entered text

6 Extracting relevant verb forms

So far, several results have been presented, concerning the task of automatically extracting all the verbs in a text. However, in the Gender and Work project, only verbs expressing what people did for a living are of interest. How well do the tools and methods used perform in the task of extracting only relevant verb forms?

The first step in exploring this issue, was to compile a gold standard containing only the interesting verb forms. The gold standard was compiled by Jonas Lindström (Department of History, Uppsala university), and lists all verbs in Per Larssons dombok that have been judged as expressing an action of work.

The manual annotation described in section 2 revealed 1 769 verb forms in
the test text. The gold standard compiled by Jonas Lindström identified only 74 verbs expressing what people did for a living. Consequently, the tool yields very low precision figures when it is used for the task of identifying only verbs expressing what people did for a living, as illustrated in table 11.

| Relevant verbs analysed as verbs | Relevant verbs not analysed as verbs | Non-verbs and irrelevant verbs analysed as verbs | Prec. | Recall |
|---------------------------------|-------------------------------------|-----------------------------------------------|-------|--------|
| Morph. analyser                 | 66                                  | 8                                            | 2652  | 0.02   | 0.89   |
| Tagger                          | 62                                  | 12                                           | 1736  | 0.03   | 0.84   |
| Morph. analyser or tagger       | 71                                  | 3                                            | 2913  | 0.02   | 0.96   |
| Morph. analyser and tagger      | 57                                  | 17                                           | 1475  | 0.04   | 0.77   |

Table 11: Recall and precision for normalised, manually entered text, when only verbs expressing what people did for a living are considered relevant

Recall remains high. A total of 12 words defined as relevant verbs in the gold standard, are analysed as non-verbs by the tagger:

- **synth** - tagged as a singular, indefinite noun

  **Context:** Johan Nilsson beroppe sig på Peer Olofson i Hagghåål som Skuthet synth hafuer och huset

  **Type of work:** **syna häst**

- **befallthe** - tagged as a singular, indefinite noun

  **Context:** Tå befallthe och Befallningzmannen välbetrodde Class Classon alla grannarne uthi byen att the broarne förfärdiga skola

  **Type of work:** **befalla någon förfärdiga broar**

- **förfärdiga** - twice tagged as a plural adjective

  **Context 1:** Tå befalltthe och Befallningzmannen välbetrodde Class Classon alla grannarne uthi byen att the broarne förfärdiga skola

  **Type of work:** **förfärdiga broar**

  **Context 2:** hänne uptagha, låta komma uthi Bättre Hampn, och sedhan förfärdiga

  **Type of work:** **bärga och laga båt**

- **lefverere** - tagged as present participle

  **Context:** att Peer i Gränöö schall lefverere hustru Ingridh i Orthale een fjerding Ströming

  **Type of work:** **ta emot strömning**
• *tiufverij* - tagged as a singular, indefinite noun, which is an appropriate analysis. This is the only case in the gold standard where a construction not comprising a verb has been identified as expressing an action of work.

  **Context:** *tiufverij på Rofvor etc*
  **Type of work:** *stjäla rovor*

• *förmana* - tagged as a proper noun

  **Context:** *Tå blev Predikanthen uthi Rassboo förordnat honom enslingen förmana till sanningenness bekennellse*
  **Type of work:** *förmana någon bekännna sanningen*

• *förbundt* - tagged as a singular, indefinite noun

  **Context:** *När hoon bemäle brände sår iempete the andra förbundt*
  **Type of work:** *förbinda sår*

• *huggo* - tagged as a singular, indefinite noun

  **Context:** *dhe huggo weedh på hanss ägor*
  **Type of work:** *hugga ved*

• *slaget* - tagged as past participle

  **Context:** *slaget holmenn*
  **Type of work:** *slå hö*

• *brukat* - tagged as past participle

  **Context:** *brukat, them androm i Stadhen till præiuditz uthi therass handell*
  **Type of work:** *använda båt för handel*

• *fråntagidt* - tagged as past participle

  **Context:** *honom hanss klädher och sjööredskap fråntagidt*
  **Type of work:** *stjäla kläder och sjööredskap*

Three of the words in the above list are not analysed as verbs by the morphological analyser either: *synth* (morphological analysis: singular, indefinite noun), *lefverere* (morphological analysis: singular, indefinite noun) and *tiufverij* (no morphological analysis).

### 7 A Graphical User Interface for automatic verb extraction

The automatic verb extraction tool was integrated in a prototypical graphical user interface, aiming at facilitating the time-consuming extraction process normally performed manually by historians.

The user interface was designed as a cgi script, using Template Toolkit and Perl, and is available for testing at [http://stp.lingfil.uu.se/~evapet/eparit.cgi](http://stp.lingfil.uu.se/~evapet/eparit.cgi)
7.1 Uploading files

At the start page, the user may upload a file from his/her computer, or type the url to a file available on the Internet. The file is to be in plain text format, preferably a text from the target period (1550-1800).

![Figure 1: The start page](image)

When the input file has been submitted, the text is processed by an underlying Perl program, and the result is displayed to the user in the form of a list of all the verbs found in the text, see figure 2.

7.2 Generating a list of verbs

The first step in generating a list of verbs from the text file submitted, is tokenisation. Tokenisation is performed using svannotate, a package for tokenisation, sentence segmentation, tagging and parsing of Swedish plain text data, developed by Filip Salomonsson (Department of Linguistics and Philology, Uppsala University). The result of the tokenisation process is a text segmented into one token per line, a blank line marking the start of a new sentence. The tokenised text is then normalised, using the normalisation rules described in chapter 4, and run through the HunPOS tagger, once again using the svannotate package.

Tokens annotated as verbs are displayed to the user in the form of an alphabetically ordered list. All verb forms belonging to the same lemma according to the SALDO morphological analyser, are displayed as one entry in the verb list. Each entry in the list consists of a lemma, a frequency figure and a list of all unique verb forms found for this lemma in the specific text. If no lemma has been found by SALDO for a word form analysed as a verb by the HunPOS tagger, this token is displayed as a single unit in the verb list, with no lemma information. If several word forms have been normalised into the same token in the normalisation step, these word forms are joined into a single entry, with the different word forms separated by a /-sign.

7.3 Verb concordances

From the list of verbs generated as a result of submitting a file to the GUI, the user may click any verb, to inspect it in context. When clicking a lemma, a concordance list is displayed, showing all the occurrences in the text of the

---

3Prior to tokenisation, a pretokenisation step is performed, removing DOS-specific characters and joining tokens where the first token is at the end of a line, followed by a hyphen.
7.4 A larger context

The user may click any of the entries in the concordance list, to view the verb in a larger context, see figure 4. The default is to display 10 sentences preceding the current sentence, and 10 sentences succeeding it. The user may choose to change the size of the context, using the drop-down menu a the top of the page.

8 Conclusions

The task of manually going through large volumes of text, searching for interesting verbs and phrases, is very time-consuming. One could assume that using natural language processing techniques to automatically display potential excerpts to the user, would reduce the time and effort needed to perform this task. In this report I have described the development of a tool that automatically retrieves the verbs from a historical text, 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 91% precision and 79% recall, using morphological tools for modern Swedish, if the text is normalised into a more modern spelling before the morphological tools are applied. (Recall can be increased to 94% if a lower precision is accepted) Any errors introduced by
Further development

The present extraction tool retrieves all verbs in a text, regardless of context and structure. However, there are verbs that are less likely to be identified as relevant, e.g. auxiliary verbs and copula verbs. Moreover, in the specific Gender and Work project, transitive verb constructions and support verb constructions, such as *göra ett köp* ("make a purchase" = "buy"), are regarded as more interesting than intransitive verbs. To retrieve complex verb constructions, the extraction tool could be supplemented with techniques for analysing syntactic structures.

Furthermore, if the list of matches were ordered according to relevance rather than alphabetically, verb constructions less likely to be of interest would be suppressed. To score the relevance of a potential excerpt, previously validated extracts could be used for machine learning.

Another issue addressed in the retrieval of excerpts is to generate a normalised form of the retrieved expression. Suggestions for normalised forms could be automatically presented, using previously validated extracts combined with machine learning techniques to predict which previous matches the current expression should be grouped with.

It would also be useful to generalise the present extraction techniques to different historical periods and text types. The extraction tool is currently based on one single text, *Per Larssons dombok* from 1638. Due to the lack of spelling rules during this period, there are however probably spelling differences between texts written by different authors, for different purposes, and in different parts of the country. Moreover, it would be interesting to see how the normalisation rules in use cope with texts from different periods covered within the Gender and Work project (i.e. 1550-1800).
Figure 4: Context displayed as a result of clicking a specific concordance in the GUI
References

[1] Ingrid Almqvist. *Precision och täckning i digitaliserat äldre källmaterial.* EparIT, 2008-11-24. Projektskiss.

[2] Lars Borin, Markus Forsberg, and Lennart Lööngren. *SALDO 1.0 (Svenskt associationslexikon version 2).* Språkbanken, Göteborgs universitet, 2008-05-07.

[3] Thorsten Brants. *Tnt - a statistical part-of-speech tagger.* Proceedings of ANLP-2000, Seattle, WA, 2000.

[4] Nils Edling, editor. *Uppländska domböcker. jämte inledning, förklaringar och register utgiven genom Nils Edling.* Uppsala, 1937.

[5] Sofia Gustafson-Capkova and Britt Hartmann. Manual of the stockholm umea corpus version 2.0. Description of the content of the SUC 2.0 distribution, including the unfinished documentation by Gunnel Källgren, December 2006.

[6] Péter Halácsy, András Kornai, and Csaba Oravecz. 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, 2007.
A Normalisation rules for the morphological analyser (in the order of application)

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

## giort -> gjort
```
$line=~s/\giort\$/\gjort\$/;
```

## voro -> vore
```
$line=~s/^voro(\$/vore$/);
```

## tuinga -> tvinga
```
$line=~s/u(\vowel)/v$1/g;
```

## hvar, hvad, wore -> var, vad, vore (in accordance with the 1906 spelling reform)
```
$line=~s/^(hv|w)\$/V/g;
```

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

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

## häfdar -> hävdar
```
$line=~s/\$vowelfd/$1vd/g;
```

## affsagt -> avsagt
```
$line=~s/\$vowelaf/$1v/g;
```

## vikja, bevijja, blijja -> vika, bevisa, bliva
```
$line=~s/ij$1/j$/;
```

## Een, saaken, sökia -> En, saken, sökia
```
$line=~s/\$vowel/1/1$/;
```

## ähr, ahntaga -> är, antaga
```
$line=~s/\$vowelh\$/a$/;
```

## utshkickadhe, utshfhöra, sågher -> utshickade, utföra, säger
```
$line=~s/\$dtfgk$/h$/;
```

## varidt -> varit (in accordance with the 1906 spelling reform)
```
$line=~s/dt\$/t$/;
```

## dömber, dömbered, benämberdh -> dömt, dömdes, benämmd
```
$line=~s/\$Nn\sae?m\$/\$/;
```

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

## försöria(s) -> försörja(s)
```
$line=~s/\s\$/a$1$/;
```

## huilen -> vilken
```
$line=~s/hui/vi$/;
```

## elliest, bevilliat -> eljest, beviljat
```
$line=~s/\$ael$/j$/;
```

## schall -> skall
```
$line=~s/sch/s$/;
```

## upsteg -> uppsteg
```
$line=~s/up(\$/p$/);
```
B Normalisation rules for the tagger

```bash
$line="s/w/v/g;
```
$line =~ s/W/V/g;
$line =~ s/qv/kv/g;
$line =~ s/Qv/Kv/g;
$line =~ s/^hv/v/g;
$line =~ s/^Hv/V/g;
$line =~ s/([aou˚ aeiy¨ a¨ oAOU˚ AEIY¨ A¨ O])ff?[uv]\([aou˚ aeiy¨ a¨ o]$/1v$2/g;
$line =~ s/j?fv/v/g;
if($line =~ /[AOU˚ AEIY¨ A¨ O][aou˚ aeiy¨ a¨ o]/){
  my $lower = lcfirst($line);
  $lower =~ s/([aou˚ aeiy¨ a¨ oAOU˚ AEIY¨ A¨ O])/$1/g;
  $line = ucfirst($lower);
}
$line =~ s/([aou˚ aeiy¨ a¨ oAOU˚ AEIY¨ A¨ O])h(\[nr\])/$1$2/g;
$line =~ s/([dtfgk])h/$1/g;
$line =~ s/lli(\[ae\])/lj$1/g;
$line =~ s/f˚ adt/f˚ att/g;
$line =~ s/F˚ adt/F˚ att/g;
$line =~ s/dt/t/g;
$line =~ s/m[bp]t/mtd/g;
$line =~ s/\[\]//g;
$line =~ s/([aou˚ aeiy¨ a¨ o])fd/$1vd/g;
$line =~ s/\([Gg]\)aff?$/$1av/g;
$line =~ s/\([Ee]\)f$//$1v/g;
$line =~ s/\([Ss]\)ch/$1k/g;
$line =~ s/\([Kk]\)i¨ o/$1¨ o/g;
$line =~ s/(gg|k)ia/$1a/g;
$line =~ s/hui/vi/g;
$line =~ s/Hui/Vi/g;
$line =~ s/ch(\[bcdfghjklmnpqrstvwxz]\))/k$1/g;
$line =~ s/\([giort]\)gjort/g;
$line =~ s/\(~[Vv])oro$/\$1ore/g;
$line =~ s/\(ft\)/ft/g;
$line =~ s/\(ss\)/s/g;
$line =~ s/\([Aa]\)f(['ft])/1v$2/g;
$line =~ s/\([Aa]\)f$/1v/g;
$line =~ s/\([Bb]\)iel/$1j¨ al/g;
$line =~ s/\([Ff]\)ier/$1j¨ ar/g;
$line =~ s/\([Hh]\)ie(\[rl\])/$1j¨ a$2/g;
$line =~ s/iemt/j¨ amt/g;
$line =~ s/Iemt/J¨ amt/g;
$line =~ s/\([Ss]\)iel/$1j¨ al/g;
$line =~ s/\(~[Tt])ien$/\$1j¨ an/g;
$line =~ s/hua/va/g;
$line =~ s/Hua/Va/g;
$line =~ s/ch(\[bcdfghjklmnpqrstvwxz]\))/k$1/g;
$line =~ s/\(giort\)/gjort/g;
$line=~s/ufu/uvu/g;
$line=~s/"ter$/dår/g;
$line=~s/"Ter$/Dår/g;
$line=~s/"te$/de/g;
$line=~s/"Te$/De/g;
$line=~s/([Ee])i$/j$/1j/g;
$line=~s/([Ss])e$/iag/g;
$line=~s/([Mm])yk[ie]t+/ycket/g;
$line=~s/"tit$/dit/g;
$line=~s/"Tit$/Dit/g;
$line=~s/"teras$/deras/g;
$line=~s/"Teras$/Deras/g;
$line=~s/"tess$/dess/g;
$line=~s/^
\([Hh]\)äller$/1eller/g;
$line=~s/^
\([Hh]\)änne$/1enne/g;
$line=~s/^
\([Bb]\)elvf$/1lev/g;
$line=~s/^
\([Åå]\)re$/1r/g;
$line=~s/^
\([Ss]\)kola$/1kulle/g;
$line=~s/^
\([Vv]\)iste$/1isste/g;
$line=~s/^
\([Ff]\)ans$/1anns/g;
$line=~s/\s/vp/up/g;
$line=~s/\s/up([p])$/upp1/g;
$line=~s/\s/up$/upp/g;
$line=~s/\s/a/g;
$line=~s/\s/oe/oe/g;