Verb Phrase Extraction in a Historical Context

Eva Pettersson[1,2], *Beáta Megyesi*[1], *Joakim Nivre*[1]

(1) Department of Linguistics and Philology, Uppsala University
(2) Swedish National Graduate School of Language Technology

firstname.lastname@lingfil.uu.se

1 Introduction

In recent years, large volumes of historical text have been made digitally available. There is however still a lack of suitable language technology tools for exploring these texts in an automated way. In this paper we present a method for automatic identification and extraction of verbs and their complements in historical text. This work has been carried out in cooperation with historians in the context of the *Gender and Work* project (GaW), where researchers are building a database with information on what men and women did for a living in the Early Modern Swedish society (approx. 1550–1800) (Ågren et al., 2011). Currently, historians are manually going through historical documents, searching for relevant text passages to store in the database. In this process, it has been noticed that working activities often are described in the form of verb phrases, such as *chop wood*, *sell fish* or *serve as a maid*. An interesting language technology challenge is thus to try to automatically extract verb phrases from historical text, and present these to the historians as a list of candidate phrases for database inclusion. Ultimately, such a tool would enable the historians to fill the GaW database with relevant phrases in a shorter period of time. In Section 2 we present our proposed method, whereas the data used in our experiments are introduced in Section 3. The approaches we use for spelling normalisation are described in Section 4. Finally, results are given in Section 5, while conclusions are drawn in Section 6.

2 Method

Our proposed method for automatic verb phrase extraction is illustrated in Figure 1, where the first step is tokenisation of the historical source text by use of standard tools. For identification of verbs and their complements, further linguistic annotation in the form of tagging and parsing is called for. Due to the absence of NLP tools adapted to historical Swedish texts, the tokenised text is first normalised to a more modern spelling, before tagging and parsing is performed. This way, NLP tools available for the modern language may be used for the linguistic analysis. For tagging, we use HunPOS (Halácsy et al., 2007) with a Swedish model based on the SUC corpus. For parsing, we use MaltParser version 1.7.2 (Nivre et al., 2006a) with a pre-trained model based on the Talbanken section of the Swedish Treebank (Nivre et al., 2006b). Finally, the annotations given by the tagger and the parser are projected back to the text in its original spelling, resulting in a tagged and parsed version of the historical text, from which the verbs and their complements are extracted.

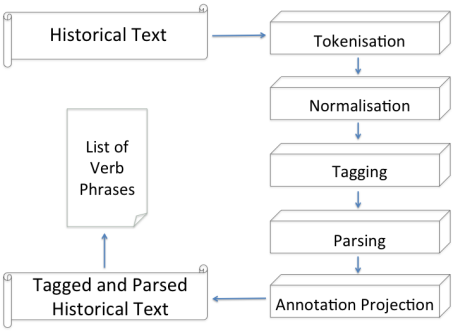

Figure 1: Method overview.

	Tokens	Types
Training corpus	28,237	7,925
Tuning corpus	2,590	1,260
Evaluation corpus	33,544	8,859

Table 1: Data used in our experiments.

3 Data

For our experiments, we make use of the Gender and Work corpus of court records and church documents from the time period 1527–1812. Three balanced subsets of this corpus are used for training, tuning and evaluation respectively, as illustrated in Table 1. In each subcorpus, the original tokens have been aligned with a manually modernised spelling of the word form in question. Furthermore, all the verbs and their complements have been manually annotated in the evaluation part of the corpus. See further Pettersson et al. (2013a) for a more detailed description of the data.

4 Spelling Normalisation

We have tried one rule-based and three data-driven approaches to spelling normalisation: dictionary-based, Levenshtein-based and SMT-based normalisation.

4.1 Rule-based Normalisation

In the rule-based approach, spelling is normalised using a set of 29 hand-crafted rules. These rules have been developed based partly on the reformed Swedish spelling introduced in 1906 (Bergman, 1995), and partly on a small sample from *Per Larssons dombok*, a court records text from 1638 (Edling, 1937). Rules based on the reformed spelling include for example the simplification of the *t* sound in spelling from *dt* to a single *t*, as in *varidt→varit* (”been”). Likewise for the *v* sound, the superfluous letters *h* or *f* were dropped, as in *hvar→var* (”was”) and *skrifva→skriva* (”write”). Examples of rules based on the court records text sample involve the substitution of letters to a phonologically similar variant, such as q→k in *qvarn→kvarn* (”mill”) and z→s in *slogz→slogs* (”were fighting”), and the deletion of repeated vowels or mute letters, as in *saak→sak* (”thing”) and *vijka→vika* (”fold”) respectively. See further Pettersson and Nivre (2011) for a more detailed description of the rule-based normalisation setting.

4.2 Dictionary-based Normalisation

In the dictionary-based approach, we use the training corpus described in Section 3 as a dictionary. Whenever a token is encountered that also occurs in the training data, the most frequent modern spelling associated with that token in the training corpus is chosen for normalisation, whereas previously unseen word forms are left unchanged.

4.3 Levenshtein-based Normalisation

In the Levenshtein-based approach, each token in the historical text is normalised by comparing the word to word forms present in a modern dictionary. The dictionary entry with the lowest Levenshtein edit distance to the original word form is chosen, provided that the distance is below a preset threshold value. If several dictionary entries share the same edit distance, the most frequent word form is chosen based on corpus data.

In our setting, we use SALDO (Borin et al., 2008) as the dictionary for Levenshtein calculations. For the frequency-based choice of a final normalisation candidate in case of a tie, we use the Stockholm Umeå corpus (SUC) (Ejerhed and Källgren, 1997). Furthermore, the training part of the GaW corpus is used as a basis for more refined Levenshtein calculations, where weights lower than 1 are assigned to frequently occurring edits observed in the training data. In addition, we try a combination of the dictionary approach and Levenshtein-based normalisation, where only tokens that are not found in the training corpus are normalised by Levenshtein comparisons. See further Pettersson et al. (2014) for a more detailed description on the Levenshtein-based normalisation approach.

4.4 SMT-based Normalisation

In the SMT-based approach, we treat spelling normalisation as a translation task, where the historical spelling is translated to a modern spelling using state-of-the-art statistical machine translation (SMT) techniques. To address changes in spelling rather than full translation of words and phrases, we perform character-based machine translation as opposed to the traditional word-based and phrase-based models.

In character-level SMT, phrases are modeled as character sequences instead of word sequences, and translation models are trained on character-aligned parallel corpora, whereas language models are trained on character N-grams. To comply with these criteria, the training and tuning corpora of token pairs mapping historical word forms to their manually modernised spelling have been adapted to a format with one token per line, with blank lines separating sentences, and with space separating the characters within each token. This will make the SMT system regard each character as a word, the full token as a sentence and the entire sentence as a paragraph. As a language model we use the SUC corpus, adapted to the same format.

The SMT engine used is Moses in its standard settings, with the GIZA++ toolkit for character alignment (Och and Ney, 2000). See further Pettersson et al. (2013b) for detailed information on the settings used in the SMT-based normalisation approach.

5 Results

Table 2 presents our results for spelling normalisation, verb identification and complement extraction, based on the evaluation corpus.

Normalisation results are given in terms of accuracy (Acc) and character error rate (CER). The baseline case shows the proportion of tokens in the original, historical text that already have a spelling identical to the modern gold standard spelling, which is true for approximately 65% of the tokens. All the proposed spelling normalisation techniques are successful in increasing the proportion of tokens with a modern spelling. The rule-based approach has a rather limited effect, increasing normalisation accuracy to approximately 73%, whereas the data-driven approaches benefit from larger datasets, yielding a normalisation accuracy of close to 93% in

	Normalisation		Verb Identification			Complement Extraction		
	Acc	CER	Pre	Rec	F-score	Pre	Rec	F-score
unnormalised	64.6	0.36	77.9	64.2	70.4	66.7	35.3	46.2
rule-based	72.7	0.28	80.0	78.0	79.0	72.4	43.4	54.3
dictionary	86.2	0.27	83.8	82.3	83.1	75.8	48.2	58.9
Levenshtein	79.4	0.22	82.0	81.8	81.9	73.7	46.0	56.6
dict+Lev	90.8	0.10	87.1	85.9	86.5	76.0	49.7	60.1
SMT	92.9	0.07	87.4	87.7	87.5	79.1	50.7	61.8
gold standard	100.0	0.00	90.1	92.3	91.2	78.1	55.2	64.7

Table 2: Results for different spelling normalisation approaches.

the SMT setting. The simplistic dictionary approach, relying solely on previously seen tokens in the training data, captures frequently occurring word forms and works surprisingly well. In fact, this method yields substantially higher normalisation results than the more sophisticated Levenshtein-based approach does. This could be partly explained by the fact that frequently occurring word forms have a high chance of being captured by the dictionary approach, whereas the Levenshtein-based approach runs the risk of consistently normalising high-frequent word forms incorrectly. There may also be old word forms that are not present in modern dictionaries and thus are out of reach for the Levenshtein-based method. Combining the dictionary method and the Levenshtein-based approach, so that only tokens that are not found in the training corpus are normalised by Levenshtein comparisons, results in higher accuracy than for any of the two methods alone, and close to the best results achieved by the SMT-based approach.

Verb identification results are presented as precision (Pre), recall (Rec) and F-score measures. In the baseline case, i.e. without any normalisation, precision is still relatively high, almost 80%. However, recall is low and we only find approximately 64% of the verbs in the evaluation corpus. With the rule-based approach, precision is still around 80%, but recall has increased significantly to 78%. Both the combined dictionary/Levenshtein approach (dict+Lev) and the SMT-based approach results in precision and recall values for verb identification that are close to the results achieved for the text in its gold standard spelling. For the manually normalised gold standard corpus, approximately 92% of the verbs are analysed as verbs by the tagger, indicating that more than spelling needs to be considered for optimal results, e.g. vocabulary and/or syntax.

Complement extraction results are given in terms of precision (Pre), recall (Rec) and F-score measures, where true positives are cases where there is a non-empty overlap between the automatically extracted complement and the gold standard complement, or where both the automatic system and the human annotator have assigned no complements at all to the verb in question. False negatives are cases where the human annotator has assigned one or more complements to a verb, whereas the system has not, including cases where the system has not even identified the verb as a verb. The results show a substantial improvement in both precision and recall with all the proposed normalisation methods. As for verb identification, SMT-based normalisation yields results close to the ones achieved for the gold standard spelling. In fact, precision is slightly higher for the text normalised by the SMT method than for the manually normalised gold standard, even though recall is still higher for the gold standard text.

6 Conclusion and Outlook

We have presented a method for linguistic analysis and information extraction based on Early Modern Swedish text, using contemporary language technology tools combined with spelling normalisation as a preprocessing step. We have shown that using this method, it is possible to automatically identify verbs and their complements in historical Swedish text with precision and recall measures close to the results achieved for manually normalised text. As future work, it would be interesting to also include syntactic and structural differences in the normalisation process. Furthermore, our work has been carried out with support from the Gender and Work project, where historians are searching for text passages describing work activities, often expressed in the form of verb phrases. Future research thus also includes automatic ranking of the extracted verb phrases, so that phrases that are more likely to describe work activities are presented at the top of the list. In addition, a user evaluation of the system is planned for.

References

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

Bergman, G. (1995). *Kortfattad svensk språkhistoria*. Prisma Magnum, Stockholm, 5th edition.

Borin, L., Forsberg, M., and Lönngren, L. (2008). Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, University of Gothenburg.

Edling, N. (1937). *Uppländska domböcker*. Almqvist & Wiksells.

Ejerhed, E. and Källgren, G. (1997). Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos - an open source trigram tagger. In *Proceedings of ACL*, pages 209–212.

Nivre, J., Hall, J., and Nilsson, J. (2006a). MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, pages 2216–2219.

Nivre, J., Nilsson, J., and Hall, J. (2006b). Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of LREC*, pages 24–26.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *ACL*, pages 440–447.

Pettersson, E., Megyesi, B., and Nivre, J. (2013a). Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of NoDaLiDa*.

Pettersson, E., Megyesi, B., and Nivre, J. (2014). A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of LaTeCH*, pages 32–41.

Pettersson, E., Megyesi, B., and Tiedemann, J. (2013b). An SMT approach to automatic annotation of historical texts. In *Workshop on Computational Historical Linguistics*.

Pettersson, E. and Nivre, J. (2011). Automatic verb extraction from historical Swedish texts. In *Proceedings of LaTeCH*, pages 87–95.