FSvReader – Exploring Old Swedish Cultural Heritage Texts

Yvonne Adesam, Malin Ahlberg, and Gerlof Bouma

Språkbanken
Department of Swedish
University of Gothenburg
firstname.lastname@gu.se

Abstract. This paper describes *FSvReader*, a tool for easier access to Old Swedish (13th–16th century) texts. Through automatic fuzzy linking of words in a text to a dictionary describing the language of the time, the reader has direct access to dictionary pop-up definitions, in spite of the large amount of morphological and spelling variation. The linked dictionary entries can also be used for simple searches in the text, highlighting possible further instances of the same entry.

1 Introduction

When exploring cultural heritage texts, different types of textual data require different types of tools for accessing the contents. For larger amounts of data, with millions or even billions of words, powerful query tools are necessary, in combination with annotations of the text, linguistic information about words and sentences in the texts. Although the added linguistic information, or annotation, may appear to be of interest only to someone exploring linguistic questions, it is in effect necessary for any user. For example, without direct access to the lemma for each word in a text, we would have to search for all the different forms of, e.g., *searching, searches, search, searched* to find mentions of people searching for something – for a language a bit more morphologically rich than English, this can be a true challenge.

For older historical text, the amount of text may not be as large, but we still require tools to glean insights from these sources as the variation in this material is typically larger than in modern material. For instance, a language's morphology may have simplified much over time (the case for, e.g., English and Swedish), and increased language standardization means less variation in vocabulary and the way words are written.

The intricacies of a language variety and of a text written in that language variety need not present serious problems for the historical linguist who is an expert of that variety. However, historical texts are also of interest to researchers from other disciplines, with different areas of expertise. In the study of a historical text, such researchers might be helped by a tool that provides additional information like dictionary definitions, and lets them quickly find related passages in the text.

We present a tool for exploring Old Swedish (Swedish: *fornsvenska*) texts (c1225–1526) – the FSvReader.[1] The tool links text words to a dictionary of Old Swedish, giving the reader quick access to dictionary definitions while reading, as well as simple search facilities of words linked to the same dictionary entry. The links are automatically created, which means that we can make large amounts of text available through the FSvReader.

2 Some Peculiarities of Historical Language: Old Swedish

Exploring historical text mostly entails close reading, and depending on the distance between the historical language and the contemporary language, this may be cumbersome. Additional knowledge is necessary to guide and assist access to the content. In the case of Old Swedish, the language has changed quite considerably in certain aspects. First, there will be word forms unfamiliar to the modern reader because of changes in morphology. As an example, the earlier Old Swedish material still uses a system with three genders and four cases, compared to modern Swedish two genders and two cases. Incidentally, these changes were already set in progress during the Old Swedish period which makes the picture even more muddled.

Secondly, some words may have disappeared from the language or may have acquired a different meaning because of semantic change. For example, Old Swedish *aptanbakka* or *affton backe* is replaced by *fladdermus* 'bat/*Chiroptera*' in modern Swedish, while the concept *bakvaþi* '(the result of) accidentally stabbing someone behind you' has altogether disappeared from the Swedish language. The Old Swedish *mot* 'meeting' developed from the noun into a preposition meaning 'towards'. As further examples of this problem we can consider jargon that is particular to (a body of) historical text, such as legal or medical terms, or standard formulations that may be assumed to have a particular meaning in some praxis described in a text. Without access to a lexicon, such cases make the text more difficult to understand. For Old Swedish, we have access to three dictionaries: the more general Söderwall and Söderwall's supplement [16] with 44,000 entries, and Schlyter [14] with 10,000 entries, focusing on law texts. These dictionaries substantially overlap in their inventory, but crucially supplement each other in their descriptions and attestations.

Thirdly and finally, some words can be difficult to understand because of spelling variation. For example, the (di)graphs *t, th, þ*[2] were used interchangeably in certain contexts, as were *d, dh, þ*, which creates a lot of uncertainty surrounding the realization of dental stops and fricatives alone.

Due to the combined sources of variation – morphological change, changes in lexical inventory, and spelling – even identifying occurrences of one and the same

[1] It has previously been briefly introduced in [2] and [1].

[2] Neither the dental fricative, nor the corresponding graph *þ*, are present in contemporary Swedish.

word can be difficult. For instance, Old Swedish **bokstaver**,[3] cf contemporary Swedish *bokstav*, 'letter', can be found in the following variants, differing both in morphology and spelling, in the editions we have digitally available: *bogstaffwa bokstaf bokstaff bokstaffua bokstaffwa bokstaffwane bokstaffwinor*[4] *bokstaffwom bokstafuom bokstafwa bokstaua bokstawa bokstawane bokstawin bokstawom*. Finding those quickly and/or automatically is a challenge.

Although a researcher working with these texts will learn to understand them, they are not easily accessible for non-expert researchers, for instance those who are just becoming acquainted with the material, let alone for the general public. The lack of an orthography is not only a problem for a modern reader of the texts, but also a major obstacle for looking up words in the dictionary. The dictionary user must both decode the form encountered in the text (requiring a passive knowledge of the spelling conventions in a material) as well as learn how the dictionary maker chose to write the base form for the expected lemma (requiring an active knowledge of the dictionary maker's spelling and/or lemmatization conventions). In some cases, the dictionary maker caters for this need, by putting alternate base forms as entries referring to the main entry describing the lemma. Yet in general, learning to use the different dictionaries in itself presents a hurdle.

As an example, one of our texts contains the multi-word unit *vkuaþins ord* 'insulting address'. Knowing a few possible character variations from just looking at the text, we can guess that *v* can also be written *u* or *o*. However, without knowing the dictionary authors' spelling conventions used for main entries – and these vary between the two main lexica available for Old Swedish – we do not know for certain where in the dictionary to look. As it turns out, our method for linking text words to dictionary entries presents the entries **oqväþins orþ**, **oqväþis ordh**, and **oqväþi** from Söderwall. Note that even though these entries come from the same dictionary, *ord* ('word') is once standardized to **þ** and once to **dh**. (See [9] for specifics on the spelling standardization in the Söderwall dictionary.) Automatic linking thus facilitates access to the historical sources, without the reader getting lost in the dictionary – even though this is a nice way to spend many hours.

3 The FSvReader

In the vein of dictionary look-up tools where you can read the text and get a pop-up dictionary definition, the FSvReader[5] contains a simple interface which allows the reader to explore the texts with the help of additional information about the words. The important difference between most such tools for modern

[3] For the rest of the paper, we will follow the convention that dictionary entries and also characters in dictionary entries are written in boldface, whereas italics refer to actual tokens and their spelling.

[4] The form *bokstaffwinor* is most likely a misspelling for *bokstaffwinom*, but we do not know where in the manuscript-edition(s)-digitalization chain this misspelling was introduced.

[5] `https://spraakbanken.gu.se/fsvreader`

	Top 1	Top 3	Top 10	# tokens
Late 13th century legal prose	66.0	89.9	95.6	1228
Mid 14th century biblical prose	60.5	78.6	83.6	7051
Mid 14th century legal prose	67.9	91.2	96.2	22107
Mid 15th century satirical prose	66.9	90.2	94.8	540
Mid 15th century fictional prose	64.9	89.3	96.2	2813
Average over texts	65.2	87.8	93.3	

Table 1. Results in percent of correctly found lexicon entries for the lexical linking method, among the top 1, 3, and 10 lexical links.

language and the FSvReader, is that we apply fuzzy matching to link words in a text to entries in the historical dictionaries for Old Swedish.

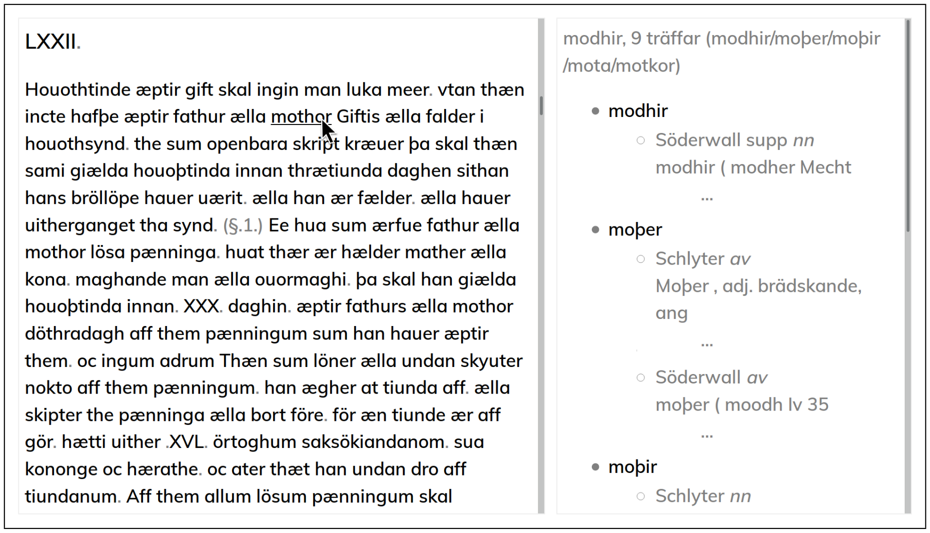
3.1 Lexical Links

Fuzzy matching allows us to link text tokens to dictionary entries without knowing the exact spelling of the lemma beforehand. The fuzzy matcher has access to a large collection of mapping rules that relate character sequences in the lemma string (the dictionary entry) to character sequences in the token string (the text word). So, for instance, we know that a dictionary **o** may be realized as a token *o* or *u*, or even *v* or *w*; a **þ** may show up as a *þ*, *t*, *d*,*th*, or *dh*; or that a lemma ending in **er**, in principle could occur in a text with a large number of endings capturing both the different suffixes of the strong masculine paradigms and their possible spellings, etc. Each of these rules has an associated weight, which encodes how likely the mapping is. The sum of the weights of all possible ways to relate a dictionary entry to a token forms a score and for a given token, we consider the top scoring dictionary entries as matches.

The weighted mapping rules are automatically induced from variants that are listed in Söderwall's dictionary. Our current fuzzy matcher knows about 171.000 rules of the kind described above. Not all rules have the convincing quality of the given examples, however, this is to be expected of rules produced by an automatic induction method. Note that a user of the end result of fuzzy matching is never confronted with these rules directly.

The fuzzy matcher is an application of spelling error correction techniques described in [6], and has been outlined in [3]. To give an idea of the quality of the resulting lemmatization, Table 1 shows the effectiveness of our method when considering just the highest scoring entry, the top 3 best entries and the top 10 entries, on five different corpora. The gold standard lemmata were annotated in the context of the morphosyntactic annotation project described in [10].

As shown in the table, the correct dictionary entry is found among the top 10 entries for more than 90% of the tokens. This average is pulled down considerably by the biblical prose, a paraphrase of the books of Moses. This text contains many proper names, which will be linked erroneously to the dictionary. In fact,



LXXII.

Houothtinde æptir gift skal ingin man luka meer. vtan thæn incte hafþe æptir fathur ælla mothor Giftis ælla falder i houothsynd. the sum openbara skript kræuer þa skal thæn sami giælda houoþtinda innan thrætiunda daghen sithan hans bröllöpe hauer uærit. ælla han ær fælder. ælla hauer uitherganget tha synd. (§.1.) Ee hua sum ærfue fathur ælla mothor lösa pænninga. huat thær ær hælder mather ælla kona. maghande man ælla ouormaghi. þa skal han giælda houoþtinda innan. XXX. daghin. æptir fathurs ælla mothor döthradagh aff them pænningum sum han hauer æptir them. oc ingum adrum Thæn sum löner ælla undan skyuter nokto aff them pænningum. han ægher at tiunda aff. ælla skipter the pænninga ælla bort före. för æn tiunde ær aff gör. hætti uither .XVL. örtoghum saksökiandanom. sua kononge oc hærathe. oc ater thæt han undan dro aff tiundanum. Aff them allum lösum pænningum skal

modhir, 9 träffar (modhir/moþer/moþir /mota/motkor)

- modhir
  - ○ Söderwall supp *nn*
    modhir (modher Mecht
    ...
- moþer
  - ○ Schlyter *av*
    Moþer , adj. brädskande, ang
    ...
  - ○ Söderwall *av*
    moþer (moodh lv 35
    ...
- moþir
  - ○ Schlyter *nn*

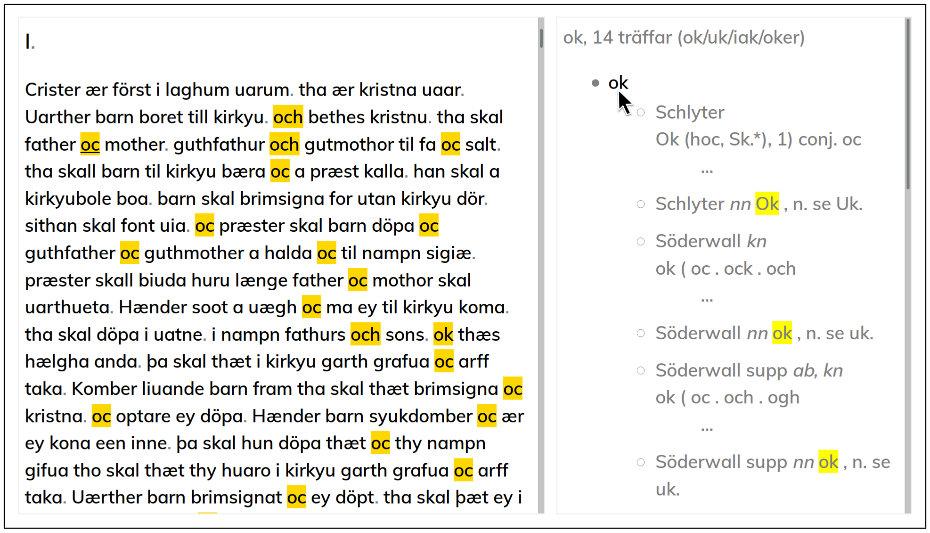
Fig. 1. The FSvReader presents the closest lexicon entries to a word in the text, here *mothor* (‘mother’) towards the end of the Church Act in *Younger Västgötalagen*

two out of three words for which we cannot find the correct match are proper names. Handling them separately would thus boost performance. Looking at the other texts, we see that considering only the top 3 matches suffices to retrieve the correct lemma in 90% of the cases.

3.2 Exploring Text with the FSvReader

The FSvReader shows a whole text to the reader. When clicking on a word in the text, the sidebar presents the closest lexicon entries to that word. An example can be found in Figure 1, where part of a law text is shown, and the word *mothor* (‘mother’ in singular, non-nominative case, underlined) on the second line of text has been clicked. Words that do not have matches are marked in grey in the text, mainly excluded items such as punctuation and numeral section headings, e.g. *4.* or *§.2.*

We have chosen to show up to three best scoring entries for each graphic token in the FSvReader, to strike a balance between finding the correct entry and not loosing the sense of overview. In addition to the best entries from the Söderwall dictionary, we also show related entries from Söderwall’s supplement and the Schlyter dictionary. By clicking on the abbreviation marker ‘. . . ’ we get access to the full lexicon entry. In the example, the first match is **modhir**, which is the correct entry for ‘mother’ in Söderwall’s supplement. The second suggestion **moþer** ‘keen/anxious’ is an example of an incorrect match. The selection is however typically small enough for a reader to quickly figure out what the correct entry is.



I.

Crister ær först i laghum uarum. tha ær kristna uaar. Uarther barn boret till kirkyu. och bethes kristnu. tha skal father oc mother. guthfathur och gutmothor til fa oc salt. tha skall barn til kirkyu bæra oc a præst kalla. han skal a kirkyubole boa. barn skal brimsigna for utan kirkyu dör. sithan skal font uia. oc præster skal barn döpa oc guthfather oc guthmother a halda oc til nampn sigiæ. præster skall biuda huru længe father oc mothor skal uarthueta. Hænder soot a uægh oc ma ey til kirkyu koma. tha skal döpa i uatne. i nampn fathurs och sons. ok thæs hælgha anda. þa skal thæt i kirkyu garth grafua oc arff taka. Komber liuande barn fram tha skal thæt brimsigna oc kristna. oc optare ey döpa. Hænder barn syukdomber oc ær ey kona een inne. þa skal hun döpa thæt oc thy nampn gifua tho skal thæt thy huaro i kirkyu garth grafua oc arff taka. Uærther barn brimsignat oc ey döpt. tha skal þæt ey i

ok, 14 träffar (ok/uk/iak/oker)

- ok
  - Schlyter
    Ok (hoc, Sk.*), 1) conj. oc
    ...
  - Schlyter *nn* Ok , n. se Uk.
  - Söderwall *kn*
    ok (oc . ock . och
    ...
  - Söderwall *nn* ok , n. se uk.
  - Söderwall supp *ab, kn*
    ok (oc . och . ogh
    ...
  - Söderwall supp *nn* ok , n. se uk.

Fig. 2. The FSvReader highlights all text words linked to the same lexicon entry, in this case *ok* (‘and’) in the start passages of the *Younger Västgötalagen*

If appropriate, we also list entries for combinations of two graphic tokens. For example, the two words *kirkmæssu dagh* are separately linked to **kirkmässa** ‘church mass’ (i.e., the anniversary of a church’s inauguration) and **dagher** ‘day’, and together to **kirkmässo dagher** ‘church mass day’, which is a separate entry in the Söderwall dictionary.

If the reader clicks on one of the lexicon entries, all words in the text linked to this entry are highlighted. In Figure 2 we can see how one instance of the word *oc* (’and’, underlined) has been clicked. Upon clicking on the (correct) entry **ok** we see several different spellings – *oc*, *och*, and *ok* – of the same word, even within the same paragraph. This is thus a simple way of finding other possible instances of the same word, regardless of their spelling.

4 Related and Future Work

There has been quite a lot of work recently on automatic methods for analysis of historical texts (see also the overview in [13]). Most approaches handle spelling variation by normalizing text words to modern spelling (e.g., [4, 5, 11, 12]), although there are approaches where standardization is done to a historical or artificial standard form (e.g., [8]). The main difference to our approach is that we do not normalize spelling, but perform lemmatization against a lexicon. The latter neutralizes inflection, and not just spelling.

Although there are many tools to analyze historical texts, there are few tools where such analysis is used to display the text and help the reader access it for close reading. A project very close in spirit to FSvReader is the digital ‘reading edition' of *Hrafnkels saga* [15],[6] which presents the Old Icelandic text with links to a dictionary, a grammar, and morphological information for words in the text. In contrast to our approach, these reading editions appear to start from existing manually annotated data, which means that the quality versus quantity balance is very different from FSvReader's.

For future work, we aim to improve the quality of the automatic dictionary linking by incorporating further linguistic knowledge, such as part-of-speech [3], which may help weed out irrelevant links. In addition, we aim to link to other resources, in particular to an onomasticon, so that we may be able to correctly recognize known proper names and also supply information about these names and their bearers. As shown in [7] for Latin, the inclusion of an onomasticon can drastically improve lemmatization quality. The treatment of lemmata that span more than two graphic tokens, and, conversely, the treatment of graphic tokens that combine several lemmata, is also planned for the future.

Further text processing, such as recognition of named entities or events, and richer semantic mark up, are very interesting, especially from a distant reading perspective. At the moment, however, we choose to concentrate our efforts for the FSvReader on the lexical level.

5 Conclusions

Reading historical text can often be difficult for someone without a detailed knowledge of the language in the text, regarding morphological and lexical differences from the modern language, as well as spelling variation. We describe a tool for exploring historical texts, which facilitates reading, by providing additional information about the words. The text words have been connected to dictionary entries with a fuzzy linking method, which gives the reader easy access to definitions, attestations, related words, etc.

The method, while also giving us erroneous links, shows us the correct entry for nine out of ten words in an Old Swedish text. In spite of the errors, this is very helpful, since manually trying to find the correct entry in the dictionary can also be difficult. We therefore hope that such a tool can help make our textual cultural heritage more accessible.

Acknowledgments

The FSvReader and its associated resources and code are freely available and accessible through `https://spraakbanken.gu.se/fsvreader`. The work of Gerlof Bouma and Yvonne Adesam was supported by the Marcus and Amalia Wallenberg foundation (MAW 2012.0146: MAÞiR).

[6] `http://ecenter.uni-tuebingen.de/hrafnkels-saga/start.html`

References

1. Adesam, Y., Ahlberg, M., Andersson, P., Borin, L., Bouma, G., Forsberg, M.: Språkteknologi för svenska språket genom tiderna. Studier i svensk språkhistoria 13, 65–87 (2016), `http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-121383`
2. Adesam, Y., Ahlberg, M., Bouma, G.: Processing spelling variation in historical text. In: Proceedings of SLTC. pp. 1–2. Lund (2012), `http://lup.lub.lu.se/record/3191709`
3. Adesam, Y., Bouma, G.: Old Swedish part-of-speech tagging between variation and external knowledge. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Berlin, Germany, August 11, 2016. Association for Computational Linguistics (2016), `https://aclweb.org/anthology/W/W16/W16-2104.pdf`
4. Baron, A.: Dealing with spelling variation in Early Modern English texts. Ph.D. thesis, Lancaster University (2011), `http://eprints.lancs.ac.uk/id/eprint/84887`
5. Bollmann, M., Petran, F., Dipper, S.: Applying rule-based normalization to different types of historical texts—an evaluation. In: Vetulani, Z., Mariani, J. (eds.) Human Language Technology Challenges for Computer Science and Linguistics. pp. 166–177. Springer International Publishing, Cham (2014)
6. Brill, E., Moore, R.C.: An improved error model for noisy channel spelling correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. pp. 286–293. ACL '00, Association for Computational Linguistics, Stroudsburg, PA, USA (2000), `https://doi.org/10.3115/1075218.1075255`
7. Budassi, M., Passarotti, M.: Nomen Omen. Enhancing the Latin morphological analyser Lemlat with an onomasticon. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Berlin, Germany, August 11, 2016. Association for Computational Linguistics (2016), `http://aclweb.org/anthology/W/W16/W16-2110.pdf`
8. Dipper, S.: Morphological and part-of-speech tagging of historical language data: A comparison. Journal for Language Technology and Computational Linguistics, Special Issue: Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities 26(2), 25–37 (2011), `http://dblp.org/rec/journals/ldvf/Dipper11a`
9. Djärv, U.: Fornsvenskans lexikala kodifiering i Söderwalls medeltidsordbok. No. 91 in Samlingar utgivna av Svenska fornskriftsällskapet. Serie 1. Svenska skrifter, Svenska fornskriftsällskapet, Uppsala (2009), `http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-26468`
10. Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O.E., Jøhndal, M.: The PROIEL treebank family: a standard for early attestations of Indo-European languages. Accepted for publication in Language Resources and Evaluation (to appear), `https://doi.org/10.1007/s10579-017-9388-5`
11. Jurish, B.: More than Words: Using Token Context to Improve Canonicalization of Historical German. Journal for Language Technology and Computational Linguistics 25(1), 23–40 (2010), `http://dblp.org/rec/journals/ldvf/Jurish10`
12. Pettersson, E.: Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. Ph.D. thesis, Uppsala University (2016), `http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-269753`
13. Piotrowski, M.: Natural Language Processing for Historical Texts. Morgan & Claypool (2012), `https://doi.org/10.2200/S00436ED1V01Y201207HLT017`

14. Schlyter, C.J.: Ordbok till Samlingen af Sweriges Gamla Lagar, Saml. af Sweriges Gamla Lagar, vol. 13. Lund (1887)
15. Schwabe, F.: Im stillen Kämmerlein. Zur digitalen Leseausgabe der ›Hrafnkels saga freysgoða‹. Magazin für digitale Editionswissenschaften 4, 21–34 (2017), https://www.mde.fau.de/?page_id=487
16. Söderwall, K.F.: Ordbok öfver svenska medeltids-språket/Ordbok över svenska medeltids-språket. Supplement. No. 27/54 in Samlingar utgivna av Svenska fornskriftsällskapet. Serie 1. Svenska skrifter., Svenska fornskriftsällskapet, Lund & Uppsala (1884–1918/1925–1973)