FACULTY OF HUMANITIES
UNIVERSITY OF COPENHAGEN

Linking corpus data to an excerpt-based historical dictionary — with > 99.9% accuracy

Tarrin Wills, Ellert Þór Jóhannsson, Simonetta Battista
Email: tarrin@hum.ku.dk, ellert@hum.ku.dk, sb@hum.ku.dk

Abstract
This paper outlines a combined automated and assisted manual process by which digital corpora can be linked to a dictionary with very high levels of accuracy (at least 99.9%) and speed — up to 4600 words per hour for previously-lemmatised XML texts and 550 words per hour for unlemmatized texts. The process combines two stages, the first of which is automated but designed to leave the resolution of ambiguity to user intervention in the second stage, which in turn assists the user by provided a range of tools at their fingertips for understanding the word and the potential lemma to which it should be linked.

Background

Texts written in Old Norse-Icelandic form a major source for the study of literature, history and culture of Viking and Medieval Scandinavia. The material consists mainly of prose texts preserved in medieval and early modern manuscripts. A great emphasis is placed in the field of Old Norse studies on the material evidence for the texts as the foundation of the discipline, namely, the manuscripts from Norway and Iceland.

The lexicography of Old Norse provides an important tool for understanding the history, literature and culture preserved in the texts. Researchers in the field using lexicographic resources expect very high levels of accuracy (at least 99.9%) and coverage (all instances of all low-frequency words, for example).

In recent years the publication of digital scholarly editions of Old Norse texts has increased manifold. Many of these texts follow the standards set by the **Medieval Nordic Text Archive (Menota)** in its published handbook (Haugen 2008). Menota "aims to preserve and publish medieval texts in digital form and to adapt and develop encoding standards necessary for this work" (http://www.menota.org).

Menota has made a large number of recent scholarly texts editions publicly available as encoded xml-files, amounting to a corpus of around 1.6 million words, most of which are within the scope of ONP's coverage, and all of which are closely based on readings of the original manuscripts of the works, the "gold standard" for ONP's corpus. Unlike ONP's traditional excerpt-based corpus, these texts provide a potential direct link between the lexicon and the manuscript page, without an intermediate edition.

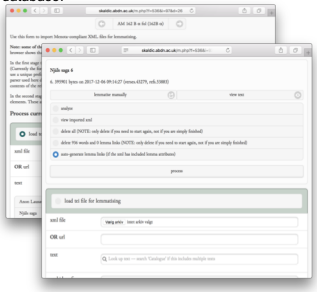
A third project, Lexicon Poeticum (LP), provides the structure and interface that allows the two others to come together. The desired outcome was that each project should benefit: Menota gains a means for automatically linking its lemmas to an authoritative external dictionary, as well as an application for assisting the manual linking process, all of which can be exported back as Menota-compatible XML; LP incorporates the remaining section of the corpus not covered by the Skaldic Project, namely the Codex Regius collection of poetry which has recently been edited according to the Menota guidelines; and ONP gains comprehensive coverage of selected manuscripts, greatly adding to its corpus, and in a format that can easily be added in the future to its own database.

Automatic lemmatising systems of Old Norse vary in accuracy, with recent published methods varying from 84% (Urban *et al.* 2014; 96% for word class) up to 92.7% accuracy for full morphosyntactic analysis (Rögnvaldsson and Helgadóttir 2011). These methods are unsuitable for a historical dictionary such as ONP: they use a highly normalised corpus and do not provide a reliable way of linking accurately from the generated lemma to a curated wordlist; and the levels of accuracy are not sufficient for the dictionary users who demand close to 100% accuracy and coverage.

Automatic lemmatising

Algorithm to link only certain matches to dictionary

Importing the XML
File is loaded, linking to text and manuscript in database.



Processing the XML
Words are imported into the database and then the automatic lemmatisation can be performed on pre-lemmatised texts:

The lemmatisation system

- A temporary reference table is built from the most recent ONP wordlist imported into the database, using the headword form, word class and noun gender to identify all unique homographs for each class/gender. This means that all potentially ambiguous homographs are ignored (e.g. the verb 'brenna' will not be used, because there are two homographs with the same word class: the weak and strong verbs).
- The database attempts to match the reference table to the lemma, word class and gender based on what was originally the lemma and me:msa attributes in the XML file.
- A link is inserted into the word table for each matching word.

Results

Menota text	Linked lemmas	Total words	Percent
Strengleikar in DG 4-7	34788	38453	90.5%
Konungs skuggsjá in AM 243 b α fol.	37299	39537	94.3%
Barlaams saga ok Jósafats in Holm perg 6	67545	76411	88.4%
Total capture	**139632**	**154401**	**90.4%**

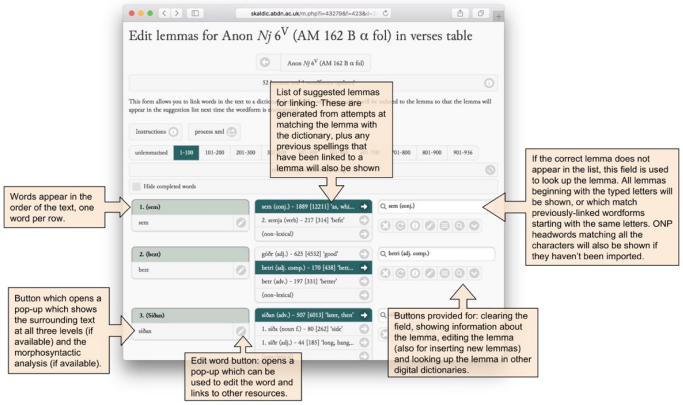
A random sample of 1000 words were checked by the authors and **no errors were introduced by this system** resulting in > 99.9% accuracy.

This system has also been tested on other corpora such as the normalised and modernised IcePaHC corpus, which does not distinguish gender of nouns, and also captures at least 80% of words.

Assisted lemmatising

Web (mobile & desktop) interface

Figure: the lemmatising form
The lemmatising form is used to assist lemmatisation, either for words not captured by the automatic method or for unlemmatized texts



User-inputted lemmatising is done by a web form which loads the words requiring lemmatising and assists in the process:

- The form uses the existing corpora of over 3,000,000 words and 300,000 distinct wordforms to find the most likely lemmas
- Additional tools in pop-ups can be used for looking up contextual information about the word and potential lemmas
- The form adapts to a range of devices including desktop computers, tablets and mobile phones

Results

Menota text	Total time (h:min)	Words lemma-tised	Time per word
Njáls saga in AM 162 B α fol.	1:12	585	7s
Njáls saga in AM 162 B θ fol.	2:52	1515	7s
Njáls saga in AM 162 B κ fol.	0:33	457	4s
Strengleikar in DG 4-7	5:20	2227	9s
Total	**9:57**	**4784**	**7s**

- The *Strengleikar* text was slightly slower because only the words not captured by the lemmatising process were lemmatised, but the inputter must still understand the word in its context. **The resulting rate for linking a previously lemmatised text was 4,600 words per hour** including the automatically linked words.
- The unlemmatized texts from *Njáls saga* was faster at this stage but included no automatically-lemmatised words. **The rate for linking and lemmatising was 550 words per hour**.

Output

XML with links to dictionary
The interface can export the XML with the links to the dictionary database and imported corpus, e.g. (fragment):

```
<w me:ref="menota:word:ends:3593121 menota:lemma:onp:51596"
lemma="maðr" me:msa="xNC gM nP cN sI">
 <choice>
  <me:facs><unclear>m<am>&bar;</am></unclear></me:facs>
  <me:dipl><unclear>m<ex>enn</ex></unclear></me:dipl>
  <me:norm>menn</me:norm>
 </choice>
</w>
<w me:ref="menota:word:ends:3593122 menota:lemma:onp:64831"
lemma="riða"
   me:msa="xVB fF mIN p3 nP vA">
 <choice>
  <me:facs><unclear>nda</unclear></me:facs>
  <me:dipl><unclear>rida</unclear></me:dipl>
  <me:norm>riða</me:norm>
 </choice>
</w>
<w me:ref="menota:word:ends:3593123 menota:lemma:onp:79587"
 lemma="til" me:msa="xAP">
 <choice>
  <me:facs><unclear>c<am>&inodotsup;</am></unclear></me:facs>
  <me:dipl><unclear>t<ex>il</ex></unclear></me:dipl>
  <me:norm>til</me:norm>
 </choice>
</w>
<w me:ref="menota:word:ends:3593124 menota:lemma:onp:89472"
 lemma="þing" me:msa="xNC gN nS cG sI">
 <choice>
  <me:facs>þingſ</me:facs>
  <me:dipl>þingſ</me:dipl>
  <me:norm>þings</me:norm>
 </choice>
</w>
```

Web interface
The web interface retains the texts and links and can be used to both view the texts and gain an overview of the use of a lemma. Additional views (not shown) give the text, manuscript images and links to other corpora.

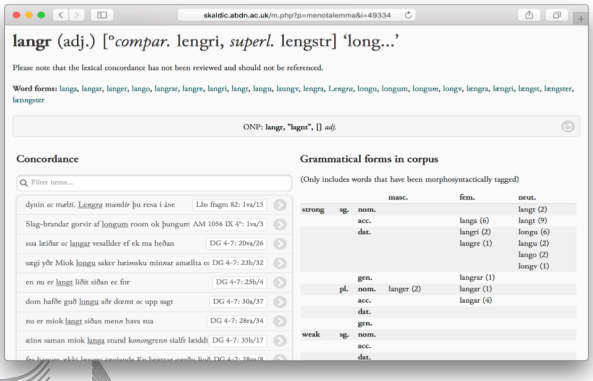
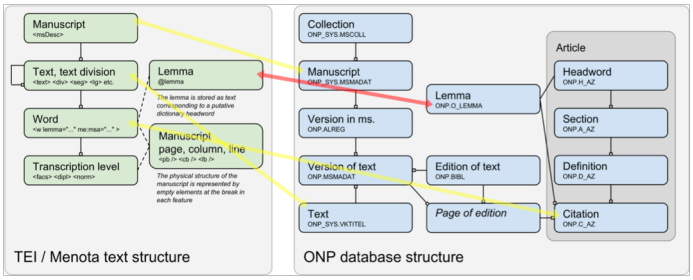


Figure: Linking TEI XML to ONP's database
The key linkage is between the textual lemma in the XML and the dictionary headword (red arrow); the process also captures contextual connections with manuscript, text and citation



European Commission | Horizon 2020 European Union funding for Research & Innovation