Abstract

This paper outlines a combined automated and assisted manual process by which digital corpus can be linked to a dictionary with very high levels of accuracy (at least 99.9%) and speed — up to 4,600 words per hour for previously-lemmatised XML texts and 550 words per hour for unlemmatised 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 providing 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. Researches 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 Augsburg ONP's coverage, and all of which are closely based on readings of the original manuscripts. For the works, the "gold standard", for ONP's, their traditional expert-based corpus, these texts provide a potential direct link between the lexicographer and the manuscript page, without an intermediate edition.

A third project, Lexicon Poeticum (LP), provides the structure and interface that allows the two other projects 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 Snæfellsnes Project, namely the Codex Regius collection of poetry which has recently been edited according to the Menota guidelines; and ONP gains a means for developing comprehensive indexes 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

Algorithm to link only certain matches to dictionary

Importing the XML File is as follows: linking to text and manuscript in database.

The lemmatisation system

A temporary reference table is built from the dictionary database and imported corpus, e.g. Menota:word:ends:3593123 menota:lemma:onp:79587.

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

<table>
<thead>
<tr>
<th>Menota text</th>
<th>Linked lemmas</th>
<th>Total words</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shingleikir in DG 4-7</td>
<td>34788</td>
<td>38453</td>
<td>90.5%</td>
</tr>
<tr>
<td>Konunga skuggas in AM 263 b fol.</td>
<td>37299</td>
<td>39637</td>
<td>94.3%</td>
</tr>
<tr>
<td>Barlaams saga ok Jóslak in Holm perg 6</td>
<td>67945</td>
<td>76461</td>
<td>88.4%</td>
</tr>
<tr>
<td>Total</td>
<td>139401</td>
<td>154401</td>
<td>90.4%</td>
</tr>
</tbody>
</table>

A random sample of 1,000 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 LatinPercorpus, which does not distinguish gender of nouns, and also captures at least 80% of words.

Output

XML with links to dictionary

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

Figure: Lexicon Poeticum (LP) 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.

Figure: Automatic lemmatising, the automatic linking of words which require linking and assists in the process.

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

Figure: Assisted lemmatising, the user inputted lemmatising is done by a web form which locates 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 lemmas 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.

Figure: Output: XML with links to dictionary is exported by the automatic method or for unlemmatised texts.