Linking Corpus Data to an Excerpt-based Historical Dictionary

Tarrin Wills, Ellert Þór Jóhannsson, Simonetta Battista
University of Copenhagen
E-mail: tarrin@hum.ku.dk, nkv950@hum.ku.dk, sb@hum.ku

Abstract

A Dictionary of Old Norse Prose (*ONP*) is a digital dictionary that derives originally from an excerpt-based index of around 750,000 citations. This paper describes recent attempts to create two-way links between the growing body of digital texts encoded using TEI XML and the dictionary's word list, which forms the basis of the published dictionary. The process involves design challenges in bringing together very different digital structures, namely the text in an XML tree structure, and the dictionary in a relational database structure. Because of the very high levels of accuracy demanded by the end-users of the dictionary (particularly researchers in Old Norse studies), the linking process can only be automated for unambiguous cases, with remaining links entered manually. The application and interface that assists this process attempts to minimize the trade-off between automation and accuracy, and adds a range of tools to assist with the human lemmatizing process. We were able to achieve linking of lemmas in 90.4% of instances where the lemma was recorded in the TEI text, with very high levels of accuracy. Where no lemma was recorded, the application allowed an Old Norse scholar to link lemmas to previously unlemmatized words at an average rate of 4-7 seconds per word.

Keywords: online dictionary, corpus material, historical lexicography, Old Norse

1 Background

Texts written in Old Norse-Icelandic (henceforward Old Norse) form a major source for the study of literature, history and culture of Viking and Medieval Scandinavia. The material consists mainly of prose narratives (sagas) as well as legal, historical and learned texts, and charters. All of these are 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, particularly the latter.

The lexicography of Old Norse provides an important tool for understanding the history, literature and culture preserved in the texts. A reliable dictionary provides a means for locating lexical indicators of sociocultural phenomena and understanding the subtleties and variations of their use. The "gold standard" of lexicography in Old Norse should provide a link not only between the lexicon and the corpus but also to its material record. 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).

A Dictionary of Old Norse Prose (*ONP*) is a dictionary project hosted at the University of Copenhagen and part of the Arnamagnæan Institute of Old Norse Manuscript Studies. The dictionary has a long history, first as an unpublished archive of dictionary citations, but later as an incomplete print edition where four out of planned 12 volumes were published. Currently *ONP* is available as an online resource at *ONP*.ku.dk. The online *ONP* brings together unedited dictionary material, material from the printed volumes, as well as more recently edited dictionary entries. The work on the dictionary continues with regular entries published online, as well as addition of new features to the online version (for a detailed overview cf. Johannsson & Battista 2014). The fundamental approach of the *ONP* is to elucidate the original medieval material while maintaining rigorous philological standards. The process involves strict textual principles and attention to orthographic details. This insistence on textual integrity has set *ONP* apart from earlier lexicographical works on Old Norse/ Icelandic. *ONP*'s corpus derives from reliable diplomatic editions that are based in turn on direct readings of the manuscripts. Each citation is linked to a published edition as well as to the manuscript which represents its primary witness. The dictionary is edited and stored as a relational database, with linked tables representing the headwords, definitions, citations, editions and manuscripts (cf. Johannsson & Battista 2016).

In recent years the publication of digital scholarly editions of Old Norse texts has increased significantly. Many of these texts follow the standards set by the Medieval Nordic Text Archive (Menota) in its published handbook (Haugen 2008), which in turn is a minor extension of the TEI XML standard, particularly the element set designed for representing primary materials. 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. LP is effectively a poetic supplement to *ONP* which is based on the digital edition of the majority of the poetic corpus made by the Skaldic Project (http://skaldic.abdn.ac.uk; Clunies Ross et al. 2007-2017). This project uses a relational database for its edition, but the textual structure was developed from a TEI model and can therefore incorporate TEI-encoded texts (Wills 2013). In addition, the data on manuscripts and prose words in the Skaldic database is based originally on *ONP*'s data, and LP uses *ONP*'s wordlist as the basis for its own. The web application developed for the Skaldic Project and LP to enter and edit data is highly extensible and forms the framework for the application described below, where the Menota texts are first incorporated into the database structure and then linked to the *ONP* wordlist.

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 lemmatizing systems of Old Norse tend to be directed towards full morphosyntactic analysis with lemma rather than as a lexicographic tool with unambiguous links between the corpus and dictionary. These automatic systems 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 & Helgadóttir 2011). These methods are unsuitable for a historical dictionary such as *ONP*: they use a highly normalized corpus compared with the manuscript-based text required by *ONP*; they do not provide a reliable way of linking accurately from the generated lemma to a curated wordlist; and the levels of accuracy, although gradually improving, are a very long way off what is required by the users of the dictionary, who demand close to 100% accuracy and coverage.

2 Method

The availability of an extensive number of digital texts that meet the textual standards of the *ONP* dictionary means that they can potentially be compared with the fragmentary digital corpus already in *ONP*. This requires aligning the corpora in a way that they can be linked and analyzed, which presents some challenges given the different nature of the data structures (relational data — linked tables; and XML — a tree structure). The two structures and points of connection between them are shown in Figure 1. *ONP*'s database encompasses a comprehensive index of all texts within its scope, and all manuscripts consulted in the editions of those texts. The Menota texts are based on transcriptions of individual manuscripts, divided into the texts which are recorded within them. TEI allows for a very complex textual structure with nested divisions, paragraphs, sentences or stanzas and lines, whereas such nesting is more difficult to encode in a relational data structure. The following process effectively "flattens" the XML structure into a series of words, but the two-way linking between the XML and database means that the details of the structure can be recovered at a later date if the database and application are updated to accommodate it.

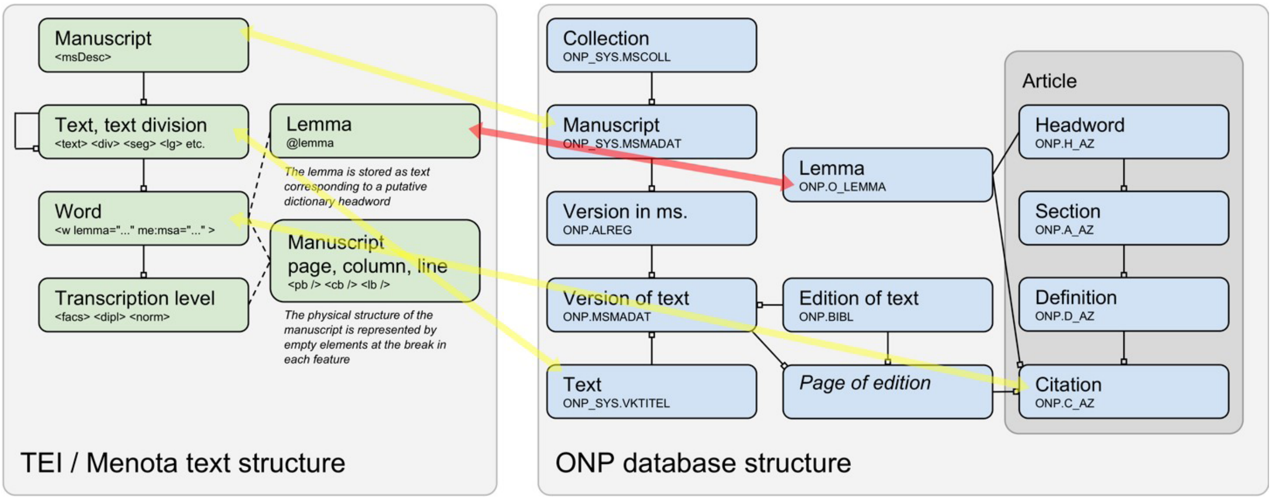


Figure 1: Menota XML structure and *ONP* database structure.

The Menota specification requires texts to be encoded with all words tagged with the <w> (word) element. This provides a functionally equivalent structure to the citation table in *ONP*, that is, an instance of a word in a text. Menota also advocates using the 'lemma' attribute of the <w> element to provide a dictionary headword for each word, providing a potential means for automating the linking process. Texts that use this attribute also encode further information in the 'me:msa' (Menota: morphosyntactic analysis) attribute, including word class, inflectional class and morphological categories.

2.1 Importing Menota XML

The Menota XML file is imported in two stages using the web application (not shown). In the first stage, the file is either uploaded or fetched via URL from the Menota catalogue and linked to both a manuscript and a text already in the database. The application reads through the file and detects all page, column and line breaks (marked up with standard TEI tags) and uses them to construct a unique identifier for all word (<w>) elements in the file, so that each word has a record of its location in the manuscript. The resulting data remains valid TEI XML and is stored on the server. In the second stage, the XML data is parsed and all word and punctuation elements (<w> and <me:punc> tags) extracted using an XPath query. The three "levels" of textual representation defined by Menota are parsed, along with the lemma attribute and morphosyntactic tokens recorded in the me:msa attribute.

The result is that the database words table is populated with the following information for each word:

- A link to the text and to the manuscript in the corresponding tables in the database
- The xml:id value for linking back to xml file
- The raw me:msa attribute string (if present) and the raw lemma attribute string (if present)
- Up to three forms of the word ("facsimile", "diplomatic" and "normalized" levels, if present)
- A number representing the position of the word in the text
- The manuscript page/folio and line number where the word begins (and column number, if relevant)
- Grammatical information from the parsed me:msa attribute: word class, strong/weak, gender, number, case, definiteness, degree of adjectives, person, tense, mood, voice, finiteness, suffix
- Punctuation for each form of the word, parsed from the following <me:punc> tag (if present)
- A link to the lemma table is undefined at this stage

This provides sufficient information for reconstructing the text, extracting the lexical context for each word, locating a word in the manuscript, and potentially automatically lemmatizing the word if sufficient information is present.

2.2 Automatic Lemmatizing

A large number of the texts in the Menota catalogue record a lemma string and morphosyntactic analysis as XML attribute values for each word. Together, the lemma value and word class provide a way of identifying and linking the word to an *ONP* dictionary headword (the 'Lemma' data types Figure 1). There is potential for ambiguity in instances where there are homographs with the same word class. Examples of this include a number of high-frequency words including the verbs *verða* (four headwords with 1,380 citations in *ONP*), *mæla* (three headwords with 948 citations) and *fá* (two headwords and 1,027 citations). In these examples one headword accounts for the overwhelming majority of instances of the word: between 96% (*mæla*) and 99.4% (*verða*) of citations, and a larger percentage of actual instances, as the citations for high-frequency words in *ONP* are not exhaustive. Researchers using the dictionary, however, require highly accurate coverage of low-frequency homographs. For example, the low-frequency homograph ^{2}*fá* (meaning to color, paint; 12 citations in *ONP*) is found in very early poetry and runic inscriptions and may be an indicator of an early date of a text. Finding all instances of this rare homograph may therefore be of interest to a historical linguist.

The following procedure is initiated by the database to avoid potentially incorrect linking of words to lemmas.

A temporary reference table is built using an SQL statement 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 process means that all potentially ambiguous homographs are ignored, e.g. the verbs *fá*, *verða* and *mæla* will not be used, because there are multiple homographs with the same word class. Despite the fact that linking to the most frequent homograph would be at least 96% accurate in these cases, it is important that all such cases are checked manually in order to capture all instances of the low-frequency words.

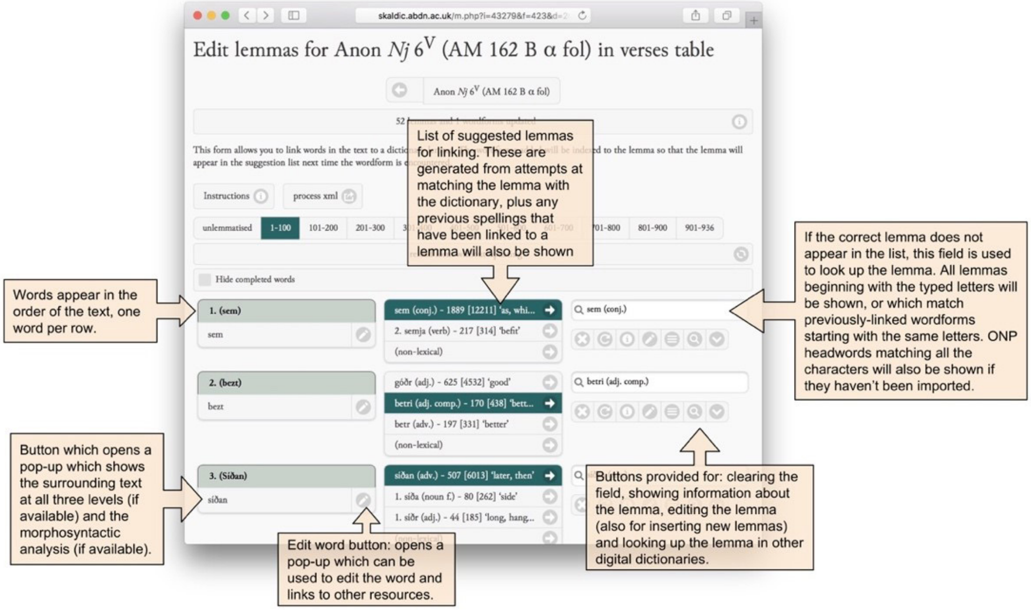
The database attempts to link the reference lemma table to the word table, using the lemma, word class and gender values based on what was originally the lemma and me:msa attributes in the XML file. Where there is a match, a link is inserted to the lemma table for each matching word in the word table. This process is initiated through the web interface and the whole process takes one or two minutes. An optional second pass attempts to match Old Norwegian variants in the lemma attributes of the remaining words and generally captures another 5% of words in Old Norwegian texts. Overall, the process captures around 90% of all words with high levels of accuracy (see Section 3 below).

The advantage of this method over one that captures more words but with lower accuracy is that the captured words do not need to be manually checked, at least in the initial stages. The remaining words comprise the ambiguous homographs mentioned above, lemmas which are not fully covered by *ONP*'s wordlist (particularly proper nouns and poetic words), or have been lemmatized in the XML according to differing practices from *ONP*'s process of determining headwords. These need to be lemmatized with human intervention, and the next stage involves using the application to assist in this process.

2.3 Assisted lemmatizing

At this stage the database contains a table of words in the manuscript text that can be used to reconstruct the text in various ways. The web application uses this table to create a web form with various features to assist in the lemmatizing of the words in the text which have not been automatically lemmatized, either because the Menota file lacked this information, or the lemmatizing process described in the previous section did not link the word to the lemma table.

The web form is shown in Figure 2 and consists of three columns, the first of which gives information about the word. In order to lemmatize the text the words must be understood in their syntactic context, and in practice it is fairly easy for the person using the form to follow the text as they scroll down the form. Where this is not possible, a pop-up can be opened which shows the word in the surrounding text, including any grammatical and lemma information that may have been previously entered.


skaldic.abdn.ac.uk/m.php?i=43279&f=423&d=2
Edit lemmas for Anon *Nj* 6^V (AM 162 B α fol) in verses table
Anon *Nj* 6^V (AM 162 B α fol)
52
This form allows you to link words in the text to a dictio
d to the lemma so that the lemma will
appear in the suggestion list next time the wordform is e
Instructions
process xml
unlemmatised
1-100
101-200
201-300
701-800
801-900
901-936
Hide completed words
1. (sem)
sem
sem (conj.) - 1889 [12211] 'as, whi...
2. semja (verb) - 217 [314] 'befit'
(non-lexical)
sem (conj.)
2. (bezt)
bezt
góðr (adj.) - 625 [4532] 'good'
betri (adj. comp.) - 170 [438] 'bett...
betr (adv.) - 197 [331] 'better'
(non-lexical)
betri (adj. comp.)
3. (Síðan)
síðan
síðan (adv.) - 507 [6013] 'later, then'
1. síða (noun f.) - 80 [262] 'side'
1. síðr (adj.) - 44 [185] 'long, hang...

List of suggested lemmas for linking. These are generated from attempts at matching the lemma with the dictionary, plus any previous spellings that have been linked to a lemma will also be shown

Words appear in the order of the text, one word per row.

If the correct lemma does not appear in the list, this field is used to look up the lemma. All lemmas beginning with the typed letters will be shown, or which match previously-linked wordforms starting with the same letters. ONP headwords matching all the characters will also be shown if they haven't been imported.

Button which opens a pop-up which shows the surrounding text at all three levels (if available) and the morphosyntactic analysis (if available).

Buttons provided for: clearing the field, showing information about the lemma, editing the lemma (also for inserting new lemmas) and looking up the lemma in other digital dictionaries.

Edit word button: opens a pop-up which can be used to edit the word and links to other resources.

Figure 2: Assisted lemmatizer form, with explanations.

The second column is a list of potential lemmas for the word. This list is generated by searching the full corpus in the database for words which are already lemmatized and match the present form in the text. The database includes around 120,000 words from the Skaldic Project, 700,000 from *ONP*'s citation list, 1,500,000 from Málföng's automatically-lemmatized corpus (cf. Rögnvaldsson and Helgadóttir 2011), as well as the words already lemmatized using this process from the Menota catalogue (200,000 words at the time of writing). The resulting matching lemmas are listed according to overall frequency. In around 90% of cases this search of the 2,500,000-word corpus produces the correct lemma as one of the options in this list.

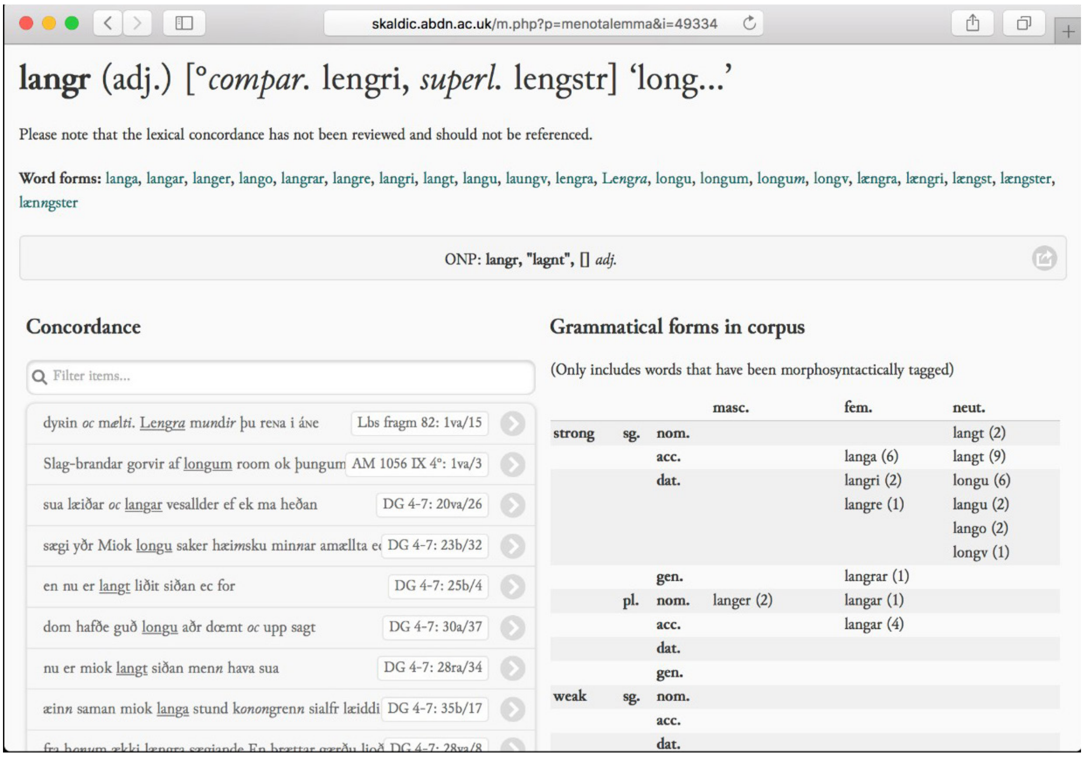
If this process does not find the correct lemma for the word, the word list can be manually searched in the third column. The results are listed in the same format as for the automatically-searched list, which includes word class, grammatical information and a gloss/definition, if available. To further assist in this process, a number of buttons provide pop-ups with further information once a potential lemma is selected, including a form for editing the lemma information itself and adding a new lemma, as well as looking up the lemma in various electronic dictionaries.

Because of the highly repetitive nature of this work there is a risk of repetitive strain injuries, particularly from using a computer mouse. The form is therefore designed to be used on a range of devices including desktop computers, tablets and phones. It is also being reviewed for optimizing for keyboard input.

Up to 100 words at a time are shown in the form and are updated when the form is submitted.

2.4 Output

The two-way linking between the XML word elements and database word table allows the inputted information to be inserted back into the XML and exported. For each word (<w>) element in the TEI XML, a “me:ref” attribute is inserted with references to the external resources in URI format (the attribute is a Menota extension). The application inserts a reference to the database key for the word in the words table (e.g. ‘menota:word:ends:3593306’ for the word *laugberkſ* on AM 162 B α fol, 1v/1) and the numeric identifier of the *ONP* headword (e.g. ‘menota:lemma:*ONP*:51275’ for *lǫgberg*). These references can be resolved as URLs using the application API.



langr (adj.) [°*compar.* lengri, *superl.* lengstr] ‘long...’

Please note that the lexical concordance has not been reviewed and should not be referenced.

Word forms: langa, langar, langer, lango, langrar, langre, langri, langt, langu, laungv, lengra, L*e*ngra, longu, longum, longu*m*, longv, længra, længri, længst, længster, læn*n*gster

ONP: **langr, "lagnt", []** *adj.*

Concordance

Filter items...

dyʀin *oc* m*æ*l*ti*. L*e*ngra m*u*nd*ir* þu reɴa i áɴe	Lbs fragm 82: 1va/15
Slag-brandar gorvir af longum room ok þungum	AM 1056 IX 4°: 1va/3
sua læiðar *oc* langar vesallder ef ek ma heðan	DG 4-7: 20va/26
sægi yðr Miok longu saker hæi*m*sku min*n*ar amællta e	DG 4-7: 23b/32
en nu er langt liðit siðan ec for	DG 4-7: 25b/4
dom hafðe guð longu aðr dœmt *oc* upp sagt	DG 4-7: 30a/37
nu er miok langt siðan men*n* hava sua	DG 4-7: 28ra/34
æin*n* saman miok langa stund k*o*n*o*ngren*n* sialfr læiddi	DG 4-7: 35b/17
fra honum ækki længra sægiande En brættar gærðu lioð	DG 4-7: 28va/8

Grammatical forms in corpus

(Only includes words that have been morphosyntactically tagged)

			masc.	fem.	neut.
strong	sg.	nom.			langt (2)
		acc.		langa (6)	langt (9)
		dat.		langri (2) langre (1)	longu (6) langu (2) lango (2) longv (1)
		gen.		langrar (1)	
	pl.	nom.	langer (2)	langar (1)	
		acc.		langar (4)	
		dat.			
		gen.			
weak	sg.	nom.			
		acc.			
		dat.			

Figure 3: Concordance and morphological forms.

If a lemma has not already been inserted, the application will populate the “lemma” attribute with the headword form from *ONP*’s wordlist. Morphosyntactic analysis can also be inputted using the assisted lemmatizing form and will be exported in Menota’s format if entered. If no analysis is entered the “me:msa” will be populated with the information that can be extracted from the word list, that is, word class and gender in the case of nouns.

The application also inserts a revision description showing which changes were made by the application and how many changes were the result of particular users. The resulting file is Menota conformant and can be imported back to the Menota Catalogue with the additional data included.

The processed and inputted data remain in the database and can be used in various ways. For example, the text can be viewed with parallel transcription levels shown in columns. Each word is linked between the three forms (shown with highlighting) and clicking/tapping on a word shows a popup with information including a link to the lemma, the word in context and grammatical information if present. The lemma is linked to further information, and the wordlist can be searched for individual lemmas. Figure 3 shows the resulting information about each lemma: the grammatical form and gloss (deriving from *ONP*); a list of word forms from the corpus; a full concordance of the word in the corpus with surrounding text, including references to the individual manuscript; and where morphosyntactic analysis has been entered, a paradigm of the lemma and its morphological forms can be reconstructed.

Additional views show the full concordance for an individual text, the text on an individual manuscript page with parallel manuscript image, and further views for the text.

3 Results

Manual updates using the lemmatizing form are logged in a separate table and each word updated this way includes a link to the editor who performed the action. We can thus extract information from the log about how long each manual operation took and how many words were captured by the automatic lemmatization process.

Table 1 shows three longer Menota XML texts which were fully or partially lemmatized and processed using the automatic lemma linking procedure outlined in 2.1 above. The total words value only includes words that were lemmatized in the XML (the *Konungs skuggsjá* text has a total of 63,895 words).

Table 1: Capture of automatic lemmatizing.

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 fol.	67545	76411	88.4%
Total	139632	154401	90.4%

A random sample of 1,000 words with surrounding text was manually checked by the authors against the *ONP* wordlist for accuracy; all (100%) were found to be correctly linked to the *ONP* lemmas based on the lemma and word class information in the Menota XML files. Two were incorrectly lemmatized in the original XML, and a very small number, although technically correct, were linked to different headwords from *ONP* due to minor differences in word class classification between *ONP* and in the original XML file. The accuracy of this process is therefore solely dependent on the accuracy of the lemmatization in the imported XML, and does not appear to introduce further errors. This means that no further systematic checking is required for the words linked by this method.

All updates using the web application are logged separately in the database and this log can be used later to analyze the processes initiated through the interface. The time taken for manual lemma linking can be calculated where there are multiple updates recorded in the database log within a limited period of time, disregarding what are clearly longer breaks between periods of lemmatizing (defined as longer than one hour). The time difference between each logged update, and the number of words altered in the relevant updates can together be used to calculate the average time per word taken to lemmatize the text. The results of this analysis are shown in Table 2.

Table 2: Time spent on manual lemmatizing.

Menota text	Total time (h:min)	Words lemmatized	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 majority of words can be linked faster than the 7s average shown in the table, but a small minority require closer investigation and in some cases the addition of lemmas to the database. These cases slow down the overall rate of lemmatizing but are necessary for the high levels of accuracy which is the aim of the system. It should also be noted that the lemmatizing process for *Strengleikar* in DG 4-7 was slightly slower than for the other texts: this is due to the fact that this text was automatically lemmatized in the first instance and only the remaining 9.5% of words were manually linked. It is faster per word to lemmatize an entire text: the user mentally parses the full sentences rather than looking at isolated words. The overall speed, however, is much faster, when 90% can be linked automatically without need for further checking.

These results are based on a single user (Wills), and the results and accuracy for other users will be investigated further.

4 Conclusion

Existing Menota texts with lemmas recorded as XML attribute values can be automatically linked to *ONP*'s headwords with high levels of capture (around 90%) and very high levels of accuracy (> 99.9%), meaning that further checking is not required. This provides a basis for supplementing the dictionary's articles and citations with direct links to manuscript text and images (see http://skaldic.abdn.ac.uk/m.php?p=*ONP*: 'other resources' tab for lemmas).

The web application provides a very fast way for a human user to lemmatize words that have not been lemmatized in XML or not linked by the automatic system, with words in unlemmatized texts able to be processed by a user familiar with Old Norse in about 7 seconds each. This means that very large TEI texts can be processed in a comparatively short time, adding greatly to the texts linked to *ONP*. The process also provides a much quicker way than has previously been achieved for inserting lemma values into the TEI text for use in the Menota archive.

The resulting data can be used by both the lexicographic and editing projects, and the different formats (XML and database) remain linked by the use of unique identifiers for each word. The data from this project, including the linked texts and concordance, are available at http://skaldic.abdn.ac.uk/m.php?p=menota.

References

Haugen, O.E. (2008). *The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources*. Version 2.0. Bergen: Medieval Nordic Text Archive. <http://www.menota.org/HB2_index.xml>

Johannsson, Ellert Thor & Simonetta Battista (2014). "A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition", in Andrea Abel & al. (eds.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 15-19 July 2014, Bolzano/Bozen, 169-179.

Johannsson, Ellert Thor & Simonetta Battista (2016). "Editing and Presenting Complex Source Material in an Online Dictionary: The Case of *ONP*", in Tinatin Margalitadze & Georg Meladze. (eds.): *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, 6-10 September 2016, Tbilisi, 117-128.

ONP = Degnbol, H., Jacobsen, B.C., Knirk, J.E., Rode, E., Sanders, C. & Helgadóttir, Þ. (eds.). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose*. *ONP* Registre (1989). *ONP* 1: a-bam (1994). *ONP* 2: ban-da (2000). *ONP* 3: de-em (2004). Copenhagen: Den Arnamagnæanske Kommission.

ONP Online. Ordbog over det norrøne prosasprog Online. Accessed at: http://*ONP*.ku.dk (20/03/2018)

Urban K., Tangherlini T.R., Vijūnas A., Broadwell P.M. (2014). Semi-Supervised Morphosyntactic Classification of Old Icelandic. In *PLOS ONE*, 9(7), e102366. <https://doi.org/10.1371/journal.pone.0102366>

Rögnvaldsson, E. and Helgadóttir, S.. 2011. Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In C. Sporleder et al. (eds.) Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series, pp. 63-76. Berlin: Springer.

Wills, T. (2015-). *Lexicon Poeticum*. Accessed at: http://lexic*ONP*oeticum.org [20/03/2018]

Wills, T. (2015). Relational data modelling of textual corpora: The Skaldic Project and its extensions. In *Literary and Linguistic Computing* 30(2), pp. 294–313. <https://doi.org/10.1093/llc/fqt045>