Integrating analog citations into an online dictionary

Ellert Thor Johannsson

A Dictionary of Old Norse Prose
University of Copenhagen
ellert@hum.ku.dk

Abstract. In this paper, I discuss A Dictionary of Old Norse Prose (ONP) and the principles behind its extensive archive of citations and how this archive has evolved to its current state as a component of an online dictionary. First, I account for the workflow involved in the digitization and further processing of paper citation slips. Then I describe how the slips were integrated with the rest of the dictionary material in a database structure, which serves as basis for both the editing system and the publication of the online dictionary. Finally, I briefly mention latest developments and planned improvements of ONP Online.

Keywords: lexicography, online dictionary, digitalization, database.

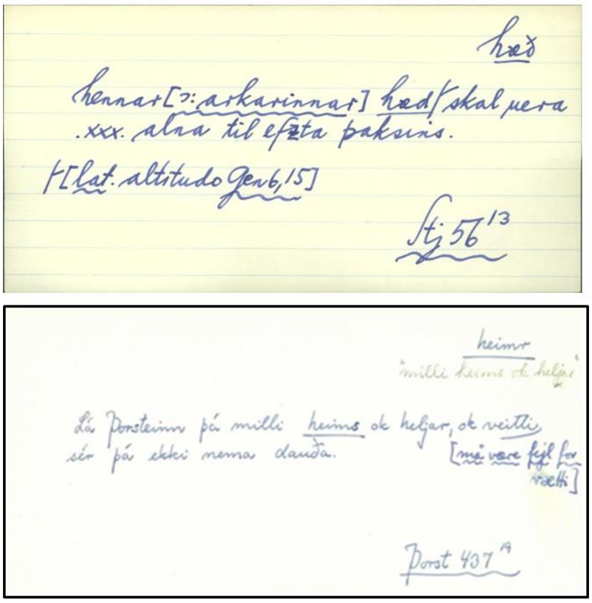
1 Background

A Dictionary of Old Norse Prose (ONP) is a dictionary project currently based at the Department of Nordic Languages and Linguistics at the University of Copenhagen. ONP records the vocabulary of prose writing in Old Norse, as transmitted in Norwegian and Icelandic manuscripts, the earliest of which date from the middle of the 12th century. The dictionary has gone through several transformations during the project's existence. After a long period of gathering lexicographic material an originally planned publication of a thirteen-volume print dictionary commenced in 1989 but was discontinued in 2004, after an index volume and three volumes of dictionary entries had been published (ONP 1-3). In 2005, the decision was made to turn ONP into a digital publication, which was to include both published and unpublished material. The digital version of the dictionary was launched in 2010 and has been available online since. ONP Online continues to grow and evolve and new features are gradually being added.

2 The citation archive and the slips

The fundamental component of ONP is its citation archive. This archive was originally created by excerpting citations from text sources, which subsequently were written on paper slips and filed in alphabetical order. The excerption work was selective, i.e. citations were chosen with the aim to exemplify the variation of the vocabulary, both in form and meaning and to give a broad overview of all the preserved genres of prose texts. In addition a few key works were excerpted completely (i.e. a slip was written out for every single word).

ONP has from its inception been guided by specific editorial principles, which involve rendering the original orthography of the source and adhering to rigorous philological standards. This entails that the citations are most commonly taken from diplomatic scholarly editions or even unpublished manuscripts and as a result the orthography is highly irregular (cf. Johannsson & Battista 2016:118-119).



hæð

hennar[ɔ: arkarinnar] hæð skal uera .xxx. alna til efsta þaksins.

[lat. altitudo Gen 6,15]

Stj 56^{13}

heimr

'milli heims ok heljar'

Lá Þorsteinn þá milli heims ok heljar, ok veitti sér þá ekki nema dauða.

[må være fejl for vætti]

Þorst 437^{19}

Fig. 1. Examples of citation slips that have some additional information, here a Latin equivalent and a comment.

Each slip is organized as follows:

- On the top to the right is the lemma or headword.
- The citation is written out with the form of the headword underlined.
- Most citations fit on a single slip, although there are some longer citations that had to be written on two slips.
- The reference siglum appears at the bottom right. This is a unique abbreviation indicates the source, both edition and the underlying manuscript, as well as page and line number.

In addition to headword, citation and reference the slips often contain various additional information, cf. Fig. 1. These can be definition suggestions, collocations and phrases, foreign equivalents (in case of prose texts translated into Old Norse) or information about variant word forms found in other versions of the same text. This additional information is often useful in the editing phase, but has traditionally been somewhat hidden, in the sense that it was only accessible through the alphabetical organization of the headwords and only then as a supplement to a particular citation and not searchable independently.

2.1 The role of citations in the print dictionary

The citation archive serves as the basis of the actual lexicographic work. When the editors started writing entries for the print edition they would read through all the citation slips associated with each headword. By analyzing the meaning of these citations, they would create the entry structure with grammatical information, division of senses and listing of phrases and collocations. Some of the supplementary information found on the slips would also be used when writing the entry, e.g. information about foreign equivalents and variant readings. The editors would also select a few demonstrative examples from the citation archive to be published in the dictionary in order to illustrate certain senses or forms that they deemed important. As part of the editing process all the citations associated with an edited headword were keyed in. However the citations used in the print edition were only about 38% of the citations that had been collected for that section of the alphabet, *a-em*.

2.2 The database and citations

It was clear that with the pace of the print publication the editing work would take many more years to complete. In order to facilitate the dissemination of the material it was decided that the dictionary should become digital and the citations themselves should be made available to the users as part of an eventual online publication. This would allow the dictionary editors to work with the citations during the editing process and at the same time give interested users access to the unpublished material.

The conclusion was to build a digital dictionary resource that would bring together several data components with the goal of an online publication. The basic component was the complete list of headwords, which already had been established in the beginning of the print publication period and existed in electronic form. Another component was all the published entries from the print publication, which also existed electronically. The third component was the citation archive, which was already partially digitized, as all the citations from *a-em* had been entered as digital text with basic markup in connection with the work on the print publication. The remaining part of the citation archive, from *en-* to the end of the alphabet only existed on paper slips. In order for the digital version to include all the citations these remaining citations, which represented about three fourths of the entire archive, needed to be digitized as well.

A database structure had been established for the dictionary in relation to the work on the print edition. This database contained all the material that existed in electronic form. The database was SQL Oracle-based and consisted of interlinked tables, which most importantly contained the list of headwords, as well as the structured dictionary entries from the published volumes and associated citations. The database also included various information about the source texts and the medieval manuscripts (see more information about the database structure in Johannsson & Battista 2016:121-122). The next step was to expand this database to include the additional unpublished citations and relevant additional information.

The optimal way to integrate the remaining citations into the database structure was not immediately apparent. It was clear that the slips needed to be digitized somehow. An OCR solution was difficult to implement as the slips were written by dozens of scribes, with handwriting of different quality and clearness. The great variation in the orthography of the texts and use of non-standard letters and symbols was also an added complication. Another possibility was to type the information found on each slip directly into the database. This idea however was abandoned quickly, as it required more resources, both time and labor, than were available. Ultimately an alternative method was selected, which involved scanning the citation slips themselves, linking them to the relevant headword in the database and in the process noting some of the information associated with each slip. This way the user would get access to the slips themselves and thereby the additional information found on the slips.

3 Digitizing and processing the slips

In order to digitize the citation slips and incorporate them into the database, the following workflow was established:

- Finalizing the list of headwords and going through the collection of citation slips, making sure that each slip was filed under the correct headword.
- Scanning all citation slips under relevant headwords in the database.
- Entering all the reference sigla from the scanned slips into the database.
- Converting outdated references to recent scholarly editions.
- Noting extra information on slips.
- Treating all variant readings.

After preparing the list of headwords, the scanning was done with a high-speed scanner where editorial assistants would work their way through the alphabet and scan the slips belonging to each headword. A simple program placed the scanned slips into the database under the relevant headword. In this way the slips were integrated into the existing data structure. However, this was only a small part of the work. In order to be able to make use of the information contained on the slips further processing was necessary.

The processing of the slips was done in several rounds. The first round involved entering the unique reference abbreviations, or sigla, into the database, along with page and line number. This was done for each of the slips. After this work was completed these old sigla were converted to new sigla, in line with the published index volume, which contained updated and more systematic abbreviations. Such conversion of references had already been done for the material published in print. In many cases, the references on the slips were obsolete, i.e. a new scholarly edition had appeared after the citation had been written up on the slip and filed in the dictionary archive. As ONP aims to cite the latest and most up-to-date available scholarly editions, such citations needed to be revised. A big part of this work was manual, as the reference to page and line needed to be checked in the new edition. The cited word itself also had to be updated if a different manuscript had been used in the new edition, a reading of a manuscript had been revised, or if the new edition followed different orthographic principles than the previous edition. Therefore, the word form itself was also entered into the database in during the conversion process.

During the conversion process, the information on the slips was also noted. This was done through a certain type of shorthand. There were several tags entered into the database with the relevant slips to reflect the type of extra information, but not the information itself. The main tags were V to indicate variant reading, U to indicate foreign equivalent word or citation, D to indicate that there was more than one citation on the slip, A to indicate that there was more than one siglum for the same citation, M to indicate a reading directly from a manuscript and K to indicate some other type of note or a comment. This tagging process enabled the sorting and processing of the slips in different ways, depending on the information found on them.

The third round involved further processing of some of these tagged slips. The biggest and most time-consuming task was the treatment of variant readings, as this work required meticulous manual look up in the relevant edition or manuscript to confirm the reading on the slip and entering the information into the database. Other tasks involved copying the slips that had more than one citation in order to link the same slip to several citation posts and confirming manuscript readings.

The whole process took about three years. The result was a database that included a digital citation archive, where every citation slip was found scanned under the relevant headword with the updated reference to a text source, along with processed variant readings and noting of various additional information. After this, the old archive of citation slips was finally fully processed and connected to the rest of the data in the database. The data could now be used to further the editing process and to create a digital version of the ONP dictionary.

4 Using the processed slips

Today all the dictionary material has been integrated and connected in the database as described above and the slips serve their twofold purpose. Firstly, they are an important tool for the editors of the dictionary as the actual lexicographic work of turning the data into structured dictionary entries continues and secondly the slips themselves are now available to the users of the dictionary and have become an important feature of the online dictionary.

4.1 The scanned slips as part of the editing process

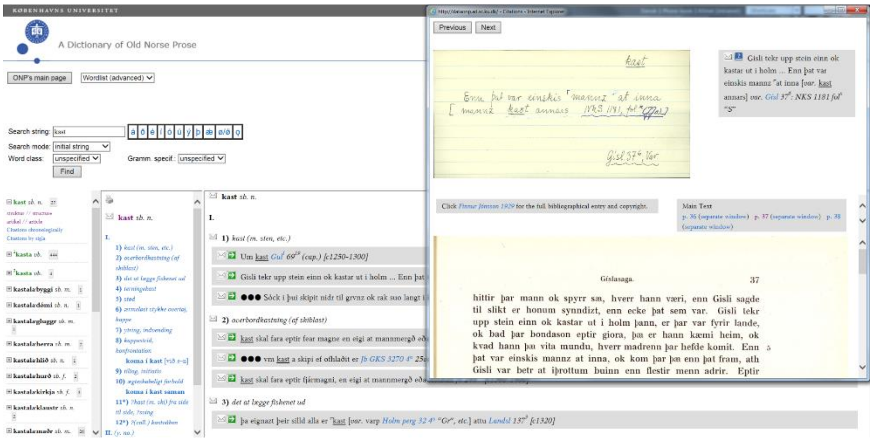
As was the case during the period of print publication, the slips play an essential part in the editing process. First the citations from relevant slips are entered into the database by dictionary assistants. The editor then has access to both the scanned slip as well as the keyed-in citation, so all eventual additional information found on the slip is immediately at hand. The editor can then easily proceed with the work on the actual entry structure by grouping the citations according to senses and highlighting some of the examples of word use.

The database has more information to facilitate its use in the editing process. The most important of those is the scanned pages of scholarly text edition. The edition and the citations are linked by the sigla in the database so the editor can look up the source of the citation slip and see the actual scanned page cited for greater context and in many cases further information from the critical apparatus of a scholarly edition. Having access to all this information has proven to increase the effectiveness of the editing procedure as the editor in most cases no longer needs to look up information in a hard copy of a book to analyze the context of a particular usage or meaning.

4.2 The scanned slips as part of an online dictionary

The scanned slips are also a prominent feature of ONP Online. When the digital version of the dictionary was finally ready to be launched in 2010 it contained all the components discussed in section 2.2, which entailed that a large portion of the data only was accessible in an unstructured form as a list of headwords with the associated scanned slips. In the period since then the editing work has progressed and as a result, the unstructured material is gradually being replaced by full-fledged dictionary entries in line with the ones found in the print publication. Nevertheless, even as unstructured groups of citation slips get replaced by dictionary entries, the slips still feature in the edited material. The user of the online dictionary is able to consult the information found on the slips within the structure of the dictionary entry.

The slips are an integrated part of ONP Online. When the user clicks on a particular citation in the entry structure a new window pops up, which shows the details of that particular citation, including both the slip and a scanned page of the edition, cf. Fig. 2.



kast sb. n.

I.

1) kast (m. sten, etc.)

Um kast Gul¹ 69¹⁹ (cap.) [c1250-1300]

Gisli tekr upp stein einn ok kastar ut i holm ... Enn þat

●●● Sóck i þui skipit nidr til grvnz ok rak suo langt

2) overbordkastning (af skiblast)

kast skal fara eptir fear magne en eigi at mannmergð eða

●●● vm kast a skipi ef ofhlaðit er Jb GKS 3270 4° 25

kast skal fara eptir fjármagni, en eigi at mannmergð eða

3) det at lægge fiskenet ud

þa eignazt þeir silld alla er ˹kast [var. varp Holm perg 32 4° "Gr", etc.] attu Landsl 137³ [c1320]

kast

Enn þat var einskis ˹mannz ˺at inna
[mannz kast annars NKS 1181, fol ˣ (Var.)

Gisl 37⁶, Var

Gisli tekr upp stein einn ok kastar ut i holm ... Enn þat var einskis mannz ˹at inna [var. kast annars] var. Gisl 37⁶: NKS 1181 fol ˣ "S"

Click Finnur Jónsson 1929 for the full bibliographical entry and copyright.

Main Text
p. 36 (separate window) p. 37 (separate window) p. 38 (separate window)

Gíslasaga. 37

hittir þar mann ok spyrr sæ, hverr hann væri, enn Gisli sagde til slikt er honum synndizt, enn ecke þat sem var. Gisli tekr upp stein einn ok kastar ut i holm þann, er þar var fyrir lande, ok bad þar bondason eptir giora, þæ er hann kæmi heim, ok kvad hann þæ vita mundu, hverr madrenn þar hefde komit. Enn 5 þat var einskis mannz at inna, ok kom þar þæ enn þat fram, ath Gisli var betr at iþrottum buinn enn flestir menn adrir. Eptir

Fig. 2. A screenshot from ONP Online, which shows a pop-up window for one particular citation from the entry structure. Here the user will find the relevant scanned slip, keyed-in citation and a scanned page from an edition.

The detailed citation view is just one of the many features of ONP Online (cf. Johannsson & Battista 2014:174-176 and 2016:124-126). The user is able to obtain greater information about each of the headwords by interacting with the data through various search features and links to supplementary material.

The scanned slips are only a feature of dictionary entries that have been edited since 2010. The slips that were used in the print publication have not been scanned and are therefore not shown in the corresponding part of the online dictionary. This creates a certain discrepancy between the entries from the first part of the alphabet and later entries. It is not inconceivable that those slips will also be scanned and incorporated into the online version at a later time, but probably not in the immediate future. ONP Online also has an additional inconsistency in the presentation of entries for verbs vs. nouns, as the verbs have only been structurally edited and the citations are not keyed in. It would be of benefit for the user if all the entries were presented in the same way, but at least for the time being, the users of the dictionary will have to accept a certain level of variance in the online entries.

Even though the slips used in the printed volumes have not yet been scanned all citations, including the ones that were left out when the examples were chosen to appear in print, have been incorporated into the relevant entry structures. The user of the online version has therefore access to all available citations. The difference between the print publication and online version of the dictionary has been discussed in some detail in Johannsson & Battista 2014 where the features of each are compared and contrasted.

5 Further developments and future improvements

ONP is in the process of redesigning the online version. This includes a different user interface with improved search capabilities, which will allow the user to take greater advantage of the interlinked data. The advanced search can be tailored in various ways, using regular expressions, since in principle each of the fields in the underlying database can form the basis of a search. This does not only include information pertaining to the slips, but also other information that has not been available to the dictionary users, such as source texts, dating of manuscripts, geographical origin or distribution between different genres.

As the editing work continues, more and more dictionary entries are provided with citations from the scanned slips that have been entered into the database. Thus, the amount of electronic, searchable text keeps increasing. In its current version ONP Online only allows search of headwords. The next logical step is to make this digital citation archive searchable and usable as a text corpus. There are some caveats, the main one being that the spelling of the citations is not normalized. This means that each variation in spelling needs to be searched separately. Nonetheless, a text search feature would allow the user to access many more examples of a particular word or combination of words, than are found under any particular headword.

When the online version of ONP was first launched more than 70% of the material was only available in raw form as scanned citation slips under a headword. Eight years later more than 70% of the material is presented in partially or fully edited entries. Concurrently with the editing of the remaining entries some efforts have been made in enhancing the ONP material with information from existing digital corpora of Old Norse texts (such as MENOTA) and modern electronic text editions (cf. Wills, Johannsson, Battista 2018). By making such text resources a feature of ONP Online additional citations can be incorporated into structured dictionary entries as well as a wider variety of searchable examples of word use will become available. In addition, various collaborative projects are being planned to provide the users of the dictionary with additional material from other related lexicographic resources, such as the Written Language Archive of Modern Icelandic *Ritmálssafn Orðabókar Háskólans* (ROH), and the new *Lexicon Poeticum project* (LP), an online dictionary of the poetic language of the medieval Norse period.

6 Conclusion

Digitizing old analog data is not always a straightforward process and even though many IT-tools can help to facilitate the procedure, the example of the ONP citation archive demonstrates how much manual work is involved in integrating such complex data into a viable digital environment. This paper shows how handwritten slips, which originally were intended to serve as the basis of a traditional paper dictionary, have been successfully incorporated into a modern online dictionary, where they form an important part of the entry structure and, in conjunction with other components, can be used in various ways. It is the editors’ hope that further development of ONP Online and the addition of new features will encourage further use of this important resource.

References

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.

LP = Lexicon Poeticum. Accessed at: http://lexiconpoeticum.org [20/10/2018]

MENOTA = Medieval Nordic Text Archive. Accessed at http://menota.org [20/10/2018]

ONP 1-3 = 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 [20/10/2018]

ROH = Ritmálssafn Orðabókar Háskólans. Stofnun Árna Magnússonar í íslenskum fræðum. Accessed at: http://ritmalssafn.arnastofnun.is [20/10/2018]

Wills, Tarrin, Ellert Thor Johannsson, Simonetta Battista. 2018. “Linking Corpus Data to an Excerpt-based Historical Dictionary”, in Čibej, Jaka, Gorjanc, Vojko, Kosem, Iztok & Krek, Simon (eds.): Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, 17-21 July 2018, Ljubljana, 979-987.