# Integrating TEI/XML Text with Semantic Lexicographic Data

Tarrin Wills[0000-0001-5360-3495], Ellert Thor Johannsson and Simonetta Battista

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

**Abstract.** Traditional excerption-based historical dictionaries often provide a very detailed semantic analysis of a high proportion of words in the corpora they cover. The Dictionary of Old Norse Prose will have analyzed and defined around 7% of all words in a 11 million word corpus, for example. Linking the semantic analysis of excerpted citations to new digital texts of the works in the corpus offers the potential to give much more detailed context for the citations in the dictionary and at the same time contextual semantic information (definitions) for a high proportion of specific words in the corpus. The task is nontrivial as it involves linking separately-formed datasets consisting of tens of thousands of tokens. This paper describes a process by which a very high proportion of citations in the dictionary are linked to individual words in new digital editions, using sorting and lexical information. The result is that users of the dictionary can view the citations in their full textual context, and readers of the text can access definitions specific to individual words in the text.

**Keywords:** Lexicography, Textual editing, TEI, XML, Databases.

## 1 Background

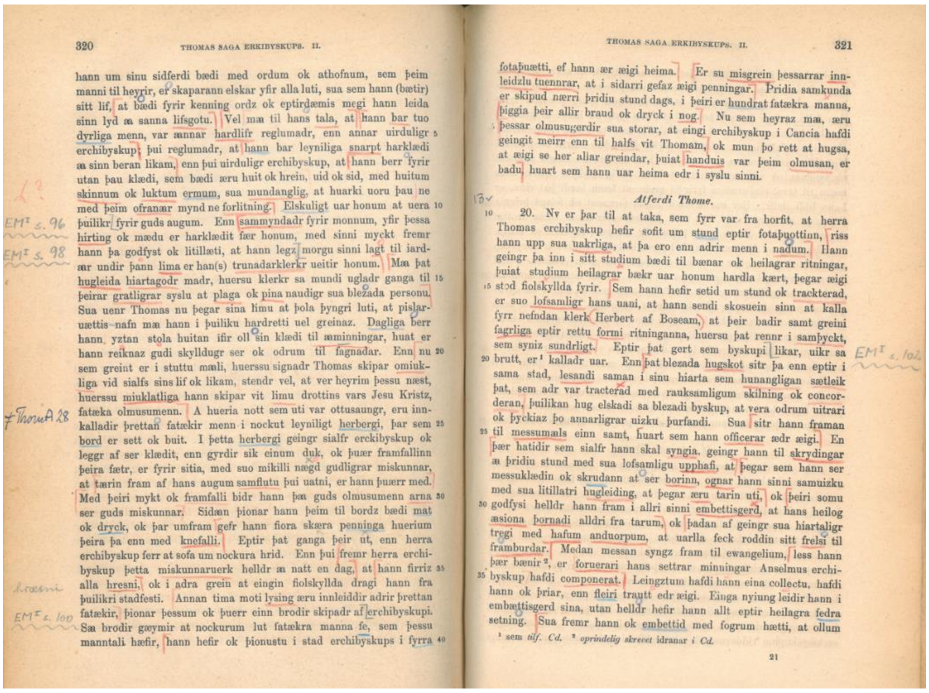
The Dictionary of Old Norse Prose (ONP: onp.ku.dk) is an extensive digital resource which links the semantic analysis of the lexicon of Old Norse with its material record (manuscripts and charters). ONP started in 1939 with a traditional process of manual excerption of words from manuscript-based editions. In this process, dictionary staff selected words from editions that were significant for lexicographic treatment. ONP works on the principle that all words are identifiable in their material context, that is, every word comes from a reading of the original manuscripts. The digital resource encodes this complex tripartite structure of words, texts and manuscripts.

An example of the excerption process for a print edition is shown in Fig. 1. Words selected for excerption were underlined and later copied onto paper slips. These have since been digitized along with metadata covering the original manuscript of the edition as well as the page and line references for each word cited.

The process of excerpting and supplementing citations has resulted in an index of around 800,000 citations from a corpus we estimate at around 11 million words. The actual number of words in original documents is much greater given the proliferation of copies of Old Norse manuscripts, but this figure represents non-repeated text. The dictionary is still in progress, with about 50% of the citations analyzed as part of finished dictionary entries. By the time of completion, the dictionary will therefore have a semantic analysis of around 7% of the words in the entire corpus. The activity of excerption means that the focus is largely on rarer words and usages, making this potentially a very rich resource for readers of these texts. Already readers can access a large number of out-of-copyright editions and view the accompanying information about each word that has been excerpted[1].

The methods and data structures for the dictionary, which began in 1939, were developed before digital corpus linguistics was possible. The dictionary's methods are based on manual excerption of words and surrounding text and are not in themselves automatically compatible with corpus-based approaches. The dictionary uses preferably diplomatic editions where orthography is not normalized, thus showing something close to the original manuscript form, and is heavily reliant on print editions. The project digitizes the citations manually, which involves some adjustment to the orthography and mark-up.

Other projects, however, have been developing manuscript-based Old Norse digital texts that belong to the same corpus covered by ONP and use a compatible manuscript-based approach.



320 THOMAS SAGA ERKIBYSKUPS. II.

hann um sinu sidferdi bædi med ordum ok athofnum, sem þeim manni til heyrir, er skaparann elskar yfir alla luti, sua sem hann (bætir) sitt lif, at bædi fyrir kenning ordz ok eptirdæmis megi hann leida sinn lyd aa sanna lifsgotu. Vel maa til hans tala, at hann bar tuo dyrliga menn, var aannar hardlifr reglumadr, enn annar uirduligr erchibyskup; þui reglumadr, at hann bar leyniliga snarpt harklædi aa sinn beran likam, enn þui uirduligr erchibyskup, at hann berr fyrir utan þau klædi, sem bædi æru huit ok hrein, uid ok sid, med huitum skinnum ok luktum ermum, sua mundanglig, at huarki uoru þau ne med þeim ofranær mynd ne forlitning. Elskuligt uar honum at uera þuilikr fyrir guds augum. Enn sammyndadr fyrir monnum, yfir þessa hirting ok mædu er harklædit fær honum, med sinni myckt fremr hann þa godfyst ok litillæti, at hann legz morgu sinni lagt til iardaar undir þann lima er han(s) trunadarklerkr ueitir honum. Maa þat hugleida hiartagodr madr, huersu klerkr sa mundi ugladr ganga til þeirar gratligrar syslu at plaga ok pina naudigr sua blezada personu. Sua uenr Thomas nu þegar sina limu at þola þyngri luti, at pislaruættis nafn maa hann i þuiliku hardretti uel greinaz. Dagliga berr hann yztan stola huitan ifir oll sin klædi til aaminningar, huat er hann reiknaz gudi skylldugr ser ok odrum til fagnadar. Enn nu sem greint er i stuttu maali, huerssu signadr Thomas skipar omiukliga vid sialfs sins lif ok likam, stendr vel, at ver heyrim þessu næst, huerssu miuklatliga hann skipar vit limu drottins vars Jesu Kristz, fatæka olmusumenn. A hueria nott sem uti var ottusaungr, eru innkalladir þrettan fatækir menn i nockut leyniligt herbergi, þar sem bord er sett ok buit. I þetta herbergi geingr sialfr erckibyskup ok leggr af ser klædit, enn gyrdir sik einum duk, ok þuær framfallinn þeira fætr, er fyrir sitia, med suo mikilli nægd gudligrar miskunnar, at tærin fram af hans augum samflutu þui uatni, er hann þuærr med. Med þeiri mykt ok framfalli bidr hann þaa guds olmusumenn arna ser guds miskunnar. Sidaan þionar hann þeim til bordz bædi mat ok dryck, ok þar umfram gefr hann fiora skæra penninga huerium þeira þa enn med knefalli. Eptir þat ganga þeir ut, enn herra erchibyskup ferr at sofa um nockura hrid. Enn þui fremr herra erchibyskup þetta miskunnaruerk helldr aa natt en dag, at hann firriz alla hresni, ok i adra grein at eingin fiolskyllda dragi hann fra þuilikri stadfesti. Annan tima moti lysing æru innleiddir adrir þrettan fatækir, þionar þessum ok þuerr einn brodir skipadr af erchibyskupi. Saa brodir gæymir at nockurum lut fatækra manna fe, sem þessu manntali hæfir, hann hefir ok þionustu i stad erchibyskups i fyrra

l?

EMI s. 96

EMI s. 98

†ThomA 28

hroesni

EMI s. 100

THOMAS SAGA ERKIBYSKUPS. II. 321

fotaþuætti, ef hann ær æigi heima. Er su misgrein þessarrar innleidzlu tuennrar, at i sidarri gefaz æigi penningar. Þridia samkunda er skipud nærri þridiu stund dags, i þeiri er hundrat fatækra manna, þiggia þeir allir braud ok dryck i nog. Nu sem heyraz maa, æru þessar olmusugerdir sua storar, at eingi erchibyskup i Cancia hafdi geingit meirr enn til halfs vit Thomam, ok mun þo rett at hugsa, at æigi se her allar greindar, þuiat handuis var þeim olmusan, er badu, huart sem hann uar heima edr i syslu sinni.

13v

*Atferdi Thome.*

20. Nv er þar til at taka, sem fyrr var fra horfit, at herra Thomas erchibyskup hefir sofit um stund eptir fotaþuottinn, riss hann upp sua uakrliga, at þa ero enn adrir menn i nadum. Hann geingr þa inn i sitt studium bædi til bænar ok heilagrar ritningar, þuiat studium heilagrar bækr uar honum hardla kært, þegar æigi stod fiolskyllda fyrir. Sem hann hefir setid um stund ok trackterad, er suo lofsamligr hans uani, at hann sendi skosuein sinn at kalla fyrr nefndan klerk Herbert af Boseam, at þeir badir samt greini fagrliga eptir rettu formi ritninganna, huersu þat rennr i samþyckt, sem syniz sundrligt. Eptir þat gert sem byskupi likar, uikr sa brutt, er[1] kalladr uar. Enn þat blezada hugskot sitr þa enn eptir i sama stad, lesandi saman i sinu hiarta sem hunangligan sætleik þat, sem adr var tracterad med rauksamligum skilning ok concorderan, þuilikan hug elskadi sa blezadi byskup, at vera odrum uitrari ok þyckiaz þo annarligrar uizku þurfandi. Sua sitr hann framan til messumaals einn samt, huart sem hann officerar ædr æigi. En þær hatidir sem sialfr hann skal syngia, geingr hann til skrydingar aa þridiu stund med sua lofsamligu upphafi, at þegar sem hann ser messuklædin ok skrudann at ser borinn, ognar hann sinni samuizku med sua litillatri hugleiding, at þegar æru tarin uti, ok þeiri somu godfysi helldr hann fram i allri sinni embettisgerd, at hans heilog aasiona þornadi alldri fra tarum, ok þadan af geingr sua hiartaligr tregi med hafum anduorpum, at uarlla feck roddin sitt frelsi til framburdar. Medan messan syngz fram til ewangelium, less hann þær bænir[2], er foruerari hans settrar minningar Anselmus erchibyskup hafdi componerat. Leingztum hafdi hann eina collectu, hafdi hann ok þriar, enn fleiri trautt edr æigi. Einga nyiung leidir hann i embættisgerd sina, utan helldr hefir hann allt eptir heilagra fedra setning. Sua fremr hann ok embettid med fogrum hætti, at ollum

EMI s. 102

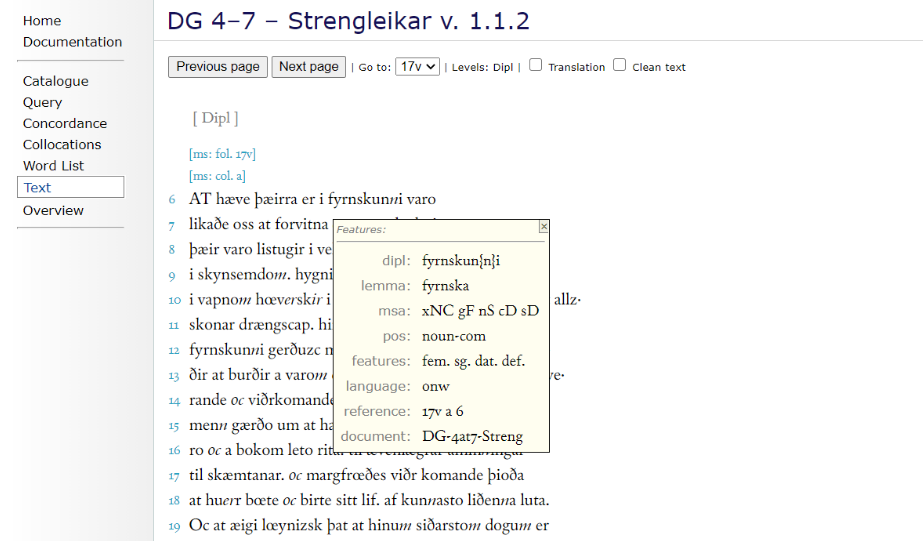
[1] sem *tilf.* Cd. [2] *oprindelig skrevet* idranar *i* Cd.

21

**Fig. 1.** Pages from the Saga of St Thomas[2] showing the words excerpted for later inclusion in the dictionary. All these words can be used to generate a contextual glossary once they are analysed.

The most extensive of these is the Menota project (menota.org[3]) which includes a catalogue of TEI/XML manuscript texts encoded according to a specified subset of TEI. Central to Menota's guidelines[4] is the principle that every word in the manuscript should be marked up with a TEI <w> (word) element. This can be extended to include up to three stages of normalization representing something very close to the manuscript orthography including abbreviations, an intermediate stage with abbreviations expanded and non-phonetic orthographic variation normalized slightly, and a fully normalized level corresponding to a kind of Classical Old Norse orthography. The markup can be further extended to include dictionary lemmas and a morphosyntactic analysis.

Menota's online catalogue[5] incorporates these features into its interface, so that users can access the information about each word (see Fig. 2). Where the text includes lemmas, the user can use them to look up the word in a dictionary, eliminating some difficulties with morphological (grammatical) or orthographic (spelling) variation. The inserted lemma, however, does not provide a concrete link to the equivalent word in a dictionary, which may use different orthography or contain different homographs.



DG 4–7 – Strengleikar v. 1.1.2

[ Dipl ]

[ms: fol. 17v]
[ms: col. a]
6 AT hæve þæirra er i fyrnskun*n*i varo
7 likaðe oss at forvitna
8 þæir varo listugir i ve
9 i skynsemdo*m*. hygni
10 i vapno*m* hœver*s*k*ir* i allz·
11 skonar drængscap. hi
12 fyrnskun*n*i gerðuzc n
13 ðir at burðir a varo*m* ve·
14 rande *oc* viðrkomand
15 men*n* gærðo um at ha
16 ro *oc* a bokom leto rit
17 til skæmtanar. *oc* margfrœðes viðr komande þioða
18 at hue*r*r bœte *oc* birte sitt lif. af kun*n*asto liðen*n*a luta.
19 Oc at æigi lœynizsk þat at hinu*m* siðarsto*m* dogu*m* er

*Features:*

dipl: fyrnskun{n}i
lemma: fyrnska
msa: xNC gF nS cD sD
pos: noun-com
features: fem. sg. dat. def.
language: onw
reference: 17v a 6
document: DG-4at7-Streng

**Fig. 2.** Information linked to digital texts in the Menota archive includes lexical and morphosyntactic information, but does not easily connect to the semantic and lexicological information provided by dictionaries.

Old Norse manuscripts are notoriously variable in their orthography and use abbreviations extensively. There is a sufficiently large discrepancy in transcription practices that it is extremely difficult to use simple string comparison to match the same text in different digital versions. This is particularly the case between earlier print editions and state-of-the-art digital editions. Even among the relatively standard Menota editions there are differences in transcription practice and in the inclusion of the diplomatic and/or normalized 'levels' of the text.

A previous paper by the authors[6] describes a fast and user-friendly workflow whereby Menota texts can be linked at the lexical level to dictionary headwords in ONP using a combination of automated and manual stages. This workflow is designed to achieve very high levels of accuracy (close to 99.9%) for the automated stages. The workflow demonstrates an interoperable method whereby TEI/XML encoded texts can be integrated and linked into relational data models such as dictionaries. These methods are designed to maintain a link between the two external data sources at the level of the word so that they can be edited and maintained separately.

Linking lemmas at the lexical level means that users can access the dictionary directly by interacting digitally with the words in the text: clicking on a word, for example, can bring up a full dictionary entry regardless of homographs, and regardless of the normalisation or lemmatisation used in the particular text edition. It also means that full concordances for a particular lemma can be generated automatically. In the reverse direction, a user of the dictionary can be directed to the exact page and line of the original document. Where images are available it means that the material source of the lexicographic data can be instantly consulted. These capabilities have already been built into the web application ONP Online at onp.ku.dk for texts that have been treated in this way.

Where this process falls short is in the semantic analysis of the word in the text: a dictionary entry can tell the user all the meanings of a word, but not what the word means in that particular context. This limitation applies even when a dictionary like ONP contains that information for a large proportion of words in any given text: for almost all texts between 5% and 15% of words have been excerpted and form part of the dictionary's analysis. In some cases close to 100% have been excerpted. The excerpted words focus on rarer lemmas and usages, meaning that there is a higher proportion still of words treated by the dictionary that might either cause difficulty or be of interest to a reader. From the potential user's point of view, the words treated by the dictionary are therefore quite likely to be a high proportion of those that an intermediate user may want further information on. The potential exists in combining the two datasets to retrieve semantic information about individual words from the manuscript text itself.

The current research builds on these resources and processes to link the words of the corpus deeper into the dictionary's semantic structure. A dictionary aims for not just a lexical but also a semantic overview of the corpus. This is done in traditional dictionaries such as ONP by excerpting relevant words from the corpus and analysing every citation excerpted, building a semantic tree of how the headword is used in the texts. Every node in that tree contains a sense and a definition of that sense, forming the structure of the dictionary entry. With such dense excerption of examples in a dictionary such as ONP, it is technically possible to link a high proportion of words in a given text to a particular semantic analysis as assigned by dictionary editors. That is, a high proportion of words can be potentially specifically linked to the individual senses and definitions of the structured dictionary entry. For the user this would mean that they can find a particular sense or usage of a word in a digital edition of a manuscript in the specific context they are reading, according to the dictionary's analysis.

## 2 Method

This goal of this project is the digital linking of individual words in a text not only to the dictionary headword but to the particular citation in the semantic tree of the dictionary entry. This is a nontrivial task. For a longer text of 100,000 words where the dictionary has excerpted some 10,000 citations, any system designed to align the two datasets must potentially deal with the Cartesian product of the two sets, namely one billion possible combinations. This processing task is simplified by using additional information to reduce the number of possible combinations, including the ordering of words and a common reference point in the lemma list. The processing is done by a system suited to the alignment of large datasets.

The references in the dictionary for citations are in almost all cases to the physical page and line of the published edition. For Menota-style TEI texts the words can normally be identified by the page and line of the manuscript version of the text. The two sets of references are in largely the same order (although this cannot always be determined for two words on the same line in the dictionary) but are not otherwise compatible. Not all words are excerpted by the dictionary, leaving no simple way of aligning and linking the two types of reference. There are nevertheless a number of principles that allow us to establish a largely automated process of linking:

1. The two datasets are in exactly the same order, despite some differences in readings. If a single word is matched between the two datasets then any word before (or after) it in the one dataset will correspond to a word before (or after) it in the other dataset.
2. All the words in one dataset (the manuscript-based editions) can be accurately linked to the dictionary headwords using an existing workflow.
3. The other dataset (dictionary citations) may be incomplete but all of the incomplete subset of words are linked to the dictionary headwords, providing the same point of reference for a subset of both corpora.
4. For lower frequency words it is very likely that the dictionary has excerpted all examples in a given text.

Therefore, if a citation in the dictionary is the only example of a particular headword excerpted from the text, and if the edition has only one word linked to the same headword then it is very likely the same word in the same context in the same text. That is, the word in the text can be linked to the corresponding citation in the dictionary (and vice-versa) in the case of most low-frequency words in the edition.

The same principle applies to smaller sections of text where the two resources are known to contain the same text. Once some words are linked, the same method can be applied to the words in both datasets that occur between the linked words.

The methodology employed here has as its first stage to identify (by database queries) lemmas that appear only once in the TEI/XML text and which also appear only once in citations from the same text for that lemma in the dictionary. Because there is little chance for ambiguity that the word in such a case corresponds to the citation in the dictionary, the word and citation can be fairly reliably linked automatically. Accuracy is around 95% and so these links require manual checking that the citation is of the same word in the same context as the word in the text.

The two datasets are currently housed on different database systems, with the imported Menota data in a MariaDB server and ONP on an Oracle system. The two systems are fairly compatible, allowing for the following process:

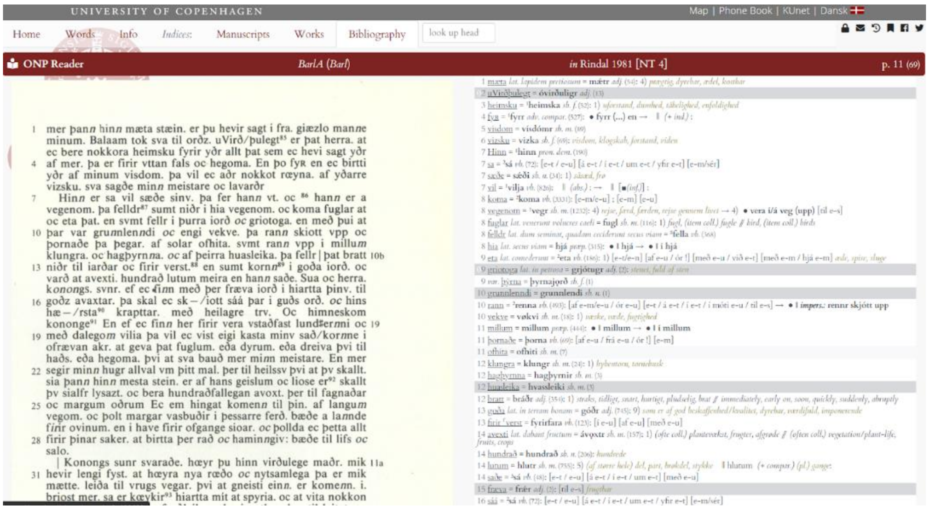
1. A subset of the word data for the text including the lemma references, ordering of the words and manuscript page/line references is exported to the ONP database.
2. A complex SQL query is run to align the words:
   a. A virtual table is built with the sections of text that are framed by the aligned words (initially the whole text), including the references to the ordering in both datasets.
   b. A virtual table is built with all the citations in ONP for the same text, including the page/line references to the edition, which forms the basis of the ordering of the cited words in ONP.
   c. These two tables are joined, and lemmas that are only found once in each section of the text among ONP's citations then form a third virtual table.
   d. A fourth virtual table links the unique words from ONP to the imported word table with the same lemmas. This table is then filtered to include lemmas that only appear once in each section of the imported text.
   e. The final table assembles then the references to the original words/citations, exporting the cross-references so that both datasets can be updated with a reference for each matched word to the corresponding word in the other dataset.
3. The exported cross-references are imported into the original data tables for each word/citation.
4. The process is then repeated so that the words matched in the previous stage can then frame the searches in more specific sections of text. In practice this is only of use for about three iterations, after which almost no further words can be matched.

A more technical description of these steps and the SQL code used can be found at https://goo.gl/ncdWAC (§9.4).

The initial links between the words and the citation index therefore provide a framework by which the same method can be applied to the smaller sections of text between the linked words. This again involves identifying lemmas unique to each section of text in both the manuscript-based edition and the print-based edition used by the dictionary. The process uses page and line references in each case to frame the extent of the section searched. Links are inserted in both data structures: in the dictionary to the word in the text, and in the text to the citation in the dictionary. This method is repeated, with decreasing gaps between the identified words, until no further automatic linking is possible. The remaining words tend to either be unlinked by the previous lemmatising processes, or are ambiguous in some way within this process and are therefore excluded.

## 3 Results

The process is extremely quick. Modern relational database management systems are optimized to align very large amounts of data and rapidly build virtual tables. The SQL query that performs the main work of this operation contains 33 lines of code and takes about 1–2 minutes to execute for each iteration on a single-processor system, allowing this process to be repeated and refined many times.



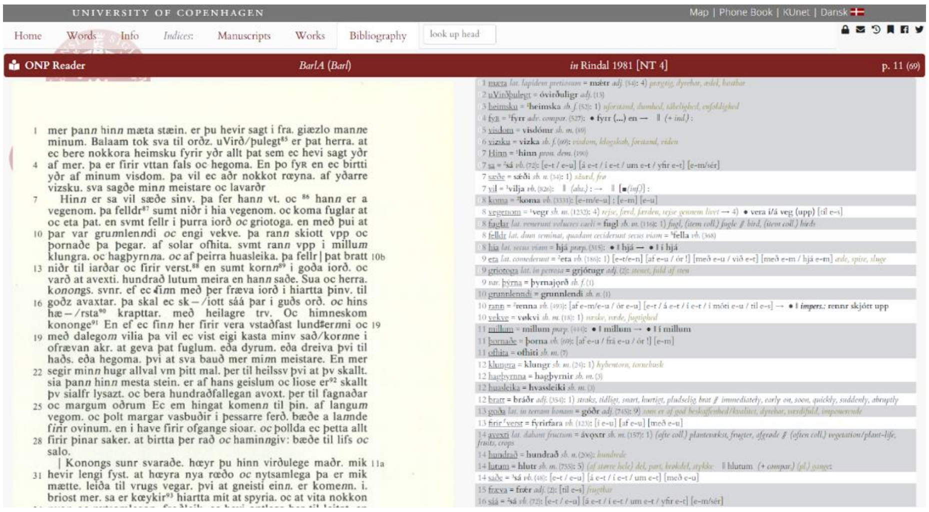
mer þann hinn mæta stæin. er þu hevir sagt i fra. giæzlo manne minum. Balaam tok sva til orðz. uVirð/þulegt[85] er þat herra. at ec bere nokkora heimsku fyrir yðr allt þat sem ec hevi sagt yðr af mer. þa er firir vttan fals oc hegoma. En þo fyʀ en ec birtti yðr af minum visdom. þa vil ec aðr nokkot rœyna. af yðarre vizsku. sva sagðe minn meistare oc lavarðr

Hinn er sa vil sæðe sinv. þa fer hann vt. oc [86] hann er a vegenom. þa felldr[87] sumt niðr i hia vegenom. oc koma fuglar at oc eta þat. en svmt fellr i þurra iorð oc griotoga. en með þui at þar var grunnlenndi oc engi vekve. þa rann skiott vpp oc þornaðe þa þegar. af solar ofhita. svmt rann vpp i millum klungra. oc hagþyrnna. oc af þeirra huasleika. þa fellr | þat bratt 10b niðr til iarðar oc firir verst.[88] en sumt kornn[89] i goða iorð. oc varð at avexti. hundrað lutum meira en hann saðe. Sua oc herra. konongs. svnr. ef ec finn með þer fræva iorð i hiartta þinv. til goðz avaxtar. þa skal ec sk–/iott sáá þar i guðs orð. oc hins hæ–/rsta[90] krapttar. með heilagre trv. Oc himneskom kononge[91] En ef ec finn her firir vera vstaðfast lundernni oc 19 með dalegom vilia þa vil ec vist eigi kasta minv sað/kornne i ofrævan akr. at geva þat fuglum. eða dyrum. eða dreiva þvi til haðs. eða hegoma. þvi at sva bauð mer minn meistare. En mer segir minn hugr allval vm þitt mal. þer til heilssv þvi at þv skallt. sia þann hinn mesta stein. er af hans geislum oc liose er[92] skallt þv sialfr lysazt. oc bera hundraðfallegan avoxt. þer til fagnaðar oc margum oðrum Ec em hingat komenn til þin. af langum vegom. oc þolt margar vasbuðir i þessarre ferð. bæðe a lannde firir ovinum. en i have firir ofgange sioar. oc þollda ec þetta allt firir þinar saker. at birtta þer rað oc haminngiv: bæðe til lifs oc salo.

| Konongs sunr svaraðe. hœyr þu hinn virðulege maðr. mik 11a hevir lengi fyst. at hœyra nya rœðo oc nytsamlega þa er mik mætte. leiða til vrugs vegar. þvi at gneisti einn. er komenn. i. briost mer. sa er kœykir[93] hiartta mit at spyria. oc at vita nokkon

1 mæta lat. lapidem pretiosum = mætr adj. (54): 4) prægtig, dyrebar, ædel, kostbar
2 uVirðþulegt = óvirðuligr adj. (13)
3 heimsku = ¹heimska sb. f. (52): 1) uforstand, dumhed, tåbelighed, enfoldighed
4 fyʀ = ¹fyrr adv. compar. (527): • fyrr (...) en → ‖ (+ ind.) :
5 visdom = vísdómr sb. m. (89)
6 vizsku = vizka sb. f. (69): visdom, klogskab, forstand, viden
7 Hinn = ¹hinn pron. dem. (190)
7 sa = ²sá vb. (72): [e-t / e-u] [á e-t / í e-t / um e-t / yfir e-t] [e-m/sér]
7 sæðe = sæði sb. n. (34): 1) sæsæd, frø
7 vil = ¹vilja vb. (826): ‖ (abs.) : → ‖ [■(inf.)] :
8 koma = ²koma vb. (3331): [e-m/e-u] : [e-m] [e-u]
8 vegenom = ¹vegr sb. m. (1232): 4) rejse, færd, færden, rejse gennem livet → 4) • vera í/á veg (upp) [til e-s]
8 fuglar lat. venerunt volucres caeli = fugl sb. m. (116): 1) fugl, (item coll.) fugle ∥ bird, (item coll.) birds
8 felldr lat. dum seminat, quaedam ceciderunt secus viam = ³fella vb. (368)
8 hia lat. secus viam = hjá prep. (315): • I hjá → • I í hjá
9 eta lat. comederunt = ²eta vb. (186): 1) [e-t/e-n] [af e-u / ór !] [með e-u / við e-t] [með e-m / hjá e-m] æde, spise, sluge
9 griotoga lat. in petrosa = grjótugr adj. (2): stenet, fuld af sten
9 var. þyrna = þyrnajǫrð sb. f. (1)
10 grunnlenndi = grunnlendi sb. n. (1)
10 rann = ²renna vb. (493): [af e-m/e-u / ór e-u] [e-t / á e-t / í e-t / í móti e-u / til e-s] → • I impers.: rennr skjótt upp
10 vekve = vøkvi sb. m. (18): 1) væske, væde, fugtighed
11 millum = millum prep. (444): • I millum → • I í millum
11 þornaðe = þorna vb. (69): [af e-u / frá e-u / ór !] [e-m]
11 ofhita = ofhiti sb. m. (7)
12 klungra = klungr sb. m. (24): 1) hybentorn, tornebusk
12 hagþyrnna = hagþyrnir sb. m. (3)
12 huasleika = hvassleiki sb. m. (3)
12 bratt = bráðr adj. (354): 1) straks, tidligt, snart, hurtigt, pludselig, brat ∥ immediately, early on, soon, quickly, suddenly, abruptly
13 goða lat. in terram bonam = góðr adj. (745): 9) som er af god beskaffenhed/kvalitet, dyrebar, værdifuld, imponerende
13 firir ˹verst = fyrirfara vb. (123): [í e-u] [af e-u] [með e-u]
14 avexti lat. dabant fructum = ávǫxtr sb. m. (157): 1) (ofte coll.) plantevækst, frugter, afgrøde ∥ (often coll.) vegetation/plant-life, fruits, crops
14 hundrað = hundrað sb. n. (206): hundrede
14 lutum = hlutr sb. m. (755): 5) (af større hele) del, part, brøkdel, stykke ‖ hlutum (+ compar.) (pl.) gange:
14 saðe = ³sá vb. (48): [e-t / e-u] [á e-t / í e-t / um e-t] [með e-u]
15 fræva = frǽr adj. (2): [til e-s] frugtbar
16 sáá = ²sá vb. (72): [e-t / e-u] [á e-t / í e-t / um e-t / yfir e-t] [e-m/sér]

**Fig. 3.** First iteration of query matches around 10% of citations (highlighted grey here) with the corresponding manuscript text.



mer þann hinn mæta stæin. er þu hevir sagt i fra. giæzlo manne minum. Balaam tok sva til orðz. uVirð/þulegt[85] er þat herra. at ec bere nokkora heimsku fyrir yðr allt þat sem ec hevi sagt yðr af mer. þa er firir vttan fals oc hegoma. En þo fyʀ en ec birtti yðr af minum visdom. þa vil ec aðr nokkot rœyna. af yðarre vizsku. sva sagðe minn meistare oc lavarðr

Hinn er sa vil sæðe sinv. þa fer hann vt. oc [86] hann er a vegenom. þa felldr[87] sumt niðr i hia vegenom. oc koma fuglar at oc eta þat. en svmt fellr i þurra iorð oc griotoga. en með þui at þar var grunnlenndi oc engi vekve. þa rann skiott vpp oc þornaðe þa þegar. af solar ofhita. svmt rann vpp i millum klungra. oc hagþyrnna. oc af þeirra huasleika. þa fellr | þat bratt 10b niðr til iarðar oc firir verst.[88] en sumt kornn[89] i goða iorð. oc varð at avexti. hundrað lutum meira en hann saðe. Sua oc herra. konongs. svnr. ef ec finn með þer fræva iorð i hiartta þinv. til goðz avaxtar. þa skal ec sk–/iott sáá þar i guðs orð. oc hins hæ–/rsta[90] krapttar. með heilagre trv. Oc himneskom kononge[91] En ef ec finn her firir vera vstaðfast lundernni oc 19 með dalegom vilia þa vil ec vist eigi kasta minv sað/kornne i ofrævan akr. at geva þat fuglum. eða dyrum. eða dreiva þvi til haðs. eða hegoma. þvi at sva bauð mer minn meistare. En mer segir minn hugr allval vm þitt mal. þer til heilssv þvi at þv skallt. sia þann hinn mesta stein. er af hans geislum oc liose er[92] skallt þv sialfr lysazt. oc bera hundraðfallegan avoxt. þer til fagnaðar oc margum oðrum Ec em hingat komenn til þin. af langum vegom. oc þolt margar vasbuðir i þessarre ferð. bæðe a lannde firir ovinum. en i have firir ofgange sioar. oc þollda ec þetta allt firir þinar saker. at birtta þer rað oc haminngiv: bæðe til lifs oc salo.

| Konongs sunr svaraðe. hœyr þu hinn virðulege maðr. mik 11a hevir lengi fyst. at hœyra nya rœðo oc nytsamlega þa er mik mætte. leiða til vrugs vegar. þvi at gneisti einn. er komenn. i. briost mer. sa er kœykir[93] hiartta mit at spyria. oc at vita nokkon

1 mæta lat. lapidem pretiosum = mætr adj. (54): 4) prægtig, dyrebar, ædel, kostbar
2 uVirðþulegt = óvirðuligr adj. (13)
3 heimsku = ¹heimska sb. f. (52): 1) uforstand, dumhed, tåbelighed, enfoldighed
4 fyʀ = ¹fyrr adv. compar. (527): • fyrr (...) en → ‖ (+ ind.) :
5 visdom = vísdómr sb. m. (89)
6 vizsku = vizka sb. f. (69): visdom, klogskab, forstand, viden
7 Hinn = ¹hinn pron. dem. (190)
7 sa = ²sá vb. (72): [e-t / e-u] [á e-t / í e-t / um e-t / yfir e-t] [e-m/sér]
7 sæðe = sæði sb. n. (34): 1) sæsæd, frø
7 vil = ¹vilja vb. (826): ‖ (abs.) : → ‖ [■(inf.)] :
8 koma = ²koma vb. (3331): [e-m/e-u] : [e-m] [e-u]
8 vegenom = ¹vegr sb. m. (1232): 4) rejse, færd, færden, rejse gennem livet → 4) • vera í/á veg (upp) [til e-s]
8 fuglar lat. venerunt volucres caeli = fugl sb. m. (116): 1) fugl, (item coll.) fugle ∥ bird, (item coll.) birds
8 felldr lat. dum seminat, quaedam ceciderunt secus viam = ³fella vb. (368)
8 hia lat. secus viam = hjá prep. (315): • I hjá → • I í hjá
9 eta lat. comederunt = ²eta vb. (186): 1) [e-t/e-n] [af e-u / ór !] [með e-u / við e-t] [með e-m / hjá e-m] æde, spise, sluge
9 griotoga lat. in petrosa = grjótugr adj. (2): stenet, fuld af sten
9 var. þyrna = þyrnajǫrð sb. f. (1)
10 grunnlenndi = grunnlendi sb. n. (1)
10 rann = ²renna vb. (493): [af e-m/e-u / ór e-u] [e-t / á e-t / í e-t / í móti e-u / til e-s] → • I impers.: rennr skjótt upp
10 vekve = vøkvi sb. m. (18): 1) væske, væde, fugtighed
11 millum = millum prep. (444): • I millum → • I í millum
11 þornaðe = þorna vb. (69): [af e-u / frá e-u / ór !] [e-m]
11 ofhita = ofhiti sb. m. (7)
12 klungra = klungr sb. m. (24): 1) hybentorn, tornebusk
12 hagþyrnna = hagþyrnir sb. m. (3)
12 huasleika = hvassleiki sb. m. (3)
12 bratt = bráðr adj. (354): 1) straks, tidligt, snart, hurtigt, pludselig, brat ∥ immediately, early on, soon, quickly, suddenly, abruptly
13 goða lat. in terram bonam = góðr adj. (745): 9) som er af god beskaffenhed/kvalitet, dyrebar, værdifuld, imponerende
13 firir ˹verst = fyrirfara vb. (123): [í e-u] [af e-u] [með e-u]
14 avexti lat. dabant fructum = ávǫxtr sb. m. (157): 1) (ofte coll.) plantevækst, frugter, afgrøde ∥ (often coll.) vegetation/plant-life, fruits, crops
14 hundrað = hundrað sb. n. (206): hundrede
14 lutum = hlutr sb. m. (755): 5) (af større hele) del, part, brøkdel, stykke ‖ hlutum (+ compar.) (pl.) gange:
14 saðe = ³sá vb. (48): [e-t / e-u] [á e-t / í e-t / um e-t] [með e-u]
15 fræva = frǽr adj. (2): [til e-s] frugtbar
16 sáá = ²sá vb. (72): [e-t / e-u] [á e-t / í e-t / um e-t / yfir e-t] [e-m/sér]

**Fig. 4.** Third iteration of the matching query showing a majority of citations linked to the corresponding word in the manuscript text (see https://onp.ku.dk/b2352-11a).

This process was tested for *Barlaams saga*, a text of some 76,000 words for which ONP has some 10,000 citations. In the first iteration 966 citations (9%) were matched. Fig. 3 shows the results with a page taken from the original non-digital edition used by the dictionary together with the matched (grey) and unmatched (white) citations excerpted from the page by the dictionary.

As the first stage has few points of alignment, the resulting matches required manual checking, aided by queries retrieving surrounding text. The second iteration, using these initial matches, produced 4,318 further matches. Subsequent iterations matched a further 430 and 23 words respectively. The result is that 56% of all citations could be matched with a very high level of accuracy by this method. Fig. 4 shows the resulting matches after three iterations of the query.

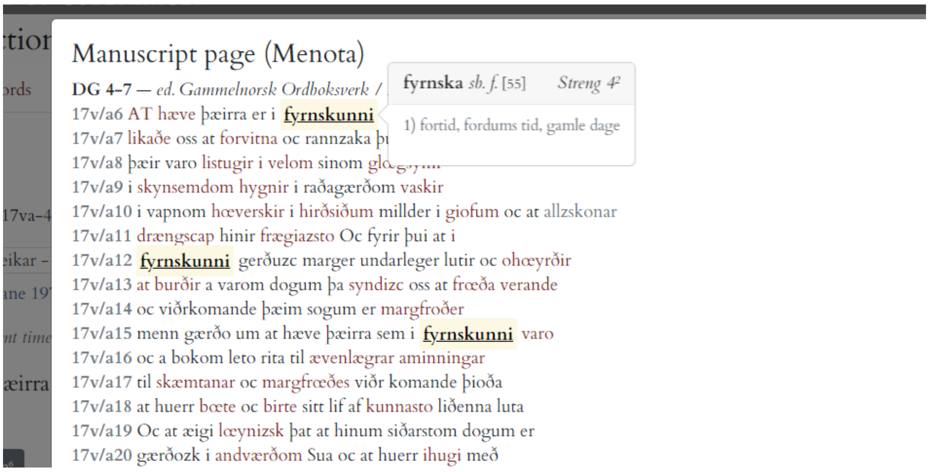
The same method has also been applied to *Strengleikar* (the Old Norse version of the *Lais* of Marie de France) in Uppsala manuscript DG 4–7 (onp.ku.dk/r10468). The method described above automatically and accurately linked 3168 of the 4065 citations in ONP to the Menota edition, representing 8% of the whole text and 78% of citations. The reasons for the much higher success rate with this text compared with *Barlaams saga* require further investigation, but it seems at this stage that a shorter text with fewer variants can be processed more successfully.

These links (as URIs and/or database keys) represent the minimal information needed to connect the words in each resource and are maintained even as the texts and the dictionary continue to be edited and developed separately.

The method is conservative and produces very few false positive matches. This requires further verification, but the authors have found very few instances where the citation is not correctly matched to the word in the manuscript text. The method, however, produces a large number of false negatives, with 20–40% of citations not matched to the corresponding words in the text. For a project such as ONP, traditional users expect extremely high levels of accuracy. Reducing the number of false negatives would likely increase the number of false positives, which would be an unacceptable trade-off in such a project: users are more likely to lose faith in the quality of the dictionary if they find information that is incorrect compared with finding information that is lacking.

The unmatched citations (false negatives) are owing to various causes. Differences in lemmatization, normalization and analytical practices mean that sometimes the same word can appear under different headwords, making them impossible to match by this method alone. In many cases a citation of a single word may appear more than once in ONP's citation index because it also records manuscript variation or other features. The alignment algorithm needs therefore to be adjusted so that it does not assume that such cases are instances of the same lemma appearing more than once, thus preventing alignment based on a unique lemma. With such refinements we expect this process to produce 80–90% matches in the future, with fewer false positives in the initial stage.

For the user the linking of words in a text to analyzed citations means that when they access the text, the individual words are not only linked to the dictionary entry, but in a good proportion of instances they are linked to the individual definition and/or phrasal-grammatical context of the word as defined in the dictionary. A user — for example a student or researcher — can pull up a section of text and click on any word to get the dictionary entry, if available. Words linked at the citation level can be highlighted to indicate that further information is linked and when clicked will show the individual definition for the word, if available, and other information that the dictionary may record about that particular citation, such as the citation slip and edition information. Users of the dictionary can find specific examples for usages and can access the full text where that usage occurs, rather than the minimal surrounding text normally provided for each citation. (For an example see Fig. 5, also at https://onp.ku.dk/c475521 and click on the Menota button. The red coloured words are linked to other citations in the dictionary, many of which have been defined.)



Manuscript page (Menota)

**DG 4–7** — *ed. Gammelnorsk Ordboksverk* /

17v/a6 AT hæve þæirra er i **fyrnskunni**
17v/a7 likaðe oss at forvitna oc rannzaka þ
17v/a8 þæir varo listugir i velom sinom glo
17v/a9 i skynsemdom hygnir i raðagærðom vaskir
17v/a10 i vapnom hœverskir i hirðsiðum millder i giofum oc at allzskonar
17v/a11 drængscap hinir frægiazsto Oc fyrir þui at i
17v/a12 **fyrnskunni** gerðuzc marger undarleger lutir oc ohœyrðir
17v/a13 at burðir a varom dogum þa syndizc oss at frœða verande
17v/a14 oc viðrkomande þæim sogum er margfroðer
17v/a15 menn gærðo um at hæve þæirra sem i **fyrnskunni** varo
17v/a16 oc a bokom leto rita til ævenlægrar aminningar
17v/a17 til skæmtanar oc margfrœðes viðr komande þioða
17v/a18 at huerr bœte oc birte sitt lif af kunnasto liðenna luta
17v/a19 Oc at æigi lœynizsk þat at hinum siðarstom dogum er
17v/a20 gærðozk i andværðom Sua oc at huerr ihugi með

**fyrnska** *sb. f.* [55] *Streng 4²*

1) fortid, fordums tid, gamle dage

**Fig. 5.** Citations are linked to the full manuscript text, which in turn can link all matched words back to the dictionary. Here a word in the text links to the specific definition for the citation in the dictionary (see https://onp.ku.dk/c185746 – ‘Menota’ button).

## 4 Discussion

The advantages of the linking method described here include the automated generation of integrated glossaries for the text edition. Such glossaries can be used as a reading aid, for those less familiar with the language, and as a language learning tool. Glosses assist in language acquisition by improving text comprehension and aiding in vocabulary acquisition[7]. This applies also to digital glosses and ‘authentic’ texts[8], such as those used in this project. While the research in this area generally focuses on living languages, it is likely that such tools would also be of benefit to students wishing to improve their reading knowledge of a historical language such as Old Norse.

The semantic analysis of significant portions of the text can be further developed if the dictionary at a later point, as is hoped, integrates a digital thesaurus. The thesaurus, when linked to particular senses in the dictionary, can be integrated into the text itself, potentially creating a semantic map of the text as a whole and helping users to find semantically similar material in the corpus. Lastly, the majority of words, those which are not analysed by the dictionary project, can be semantically analysed according to statistical or other digital methods so that the particular meanings of non-manually analysed words can potentially be predicted from those in similar contexts.

The method applied here can be potentially applied for any project where a subset of words needs to be realigned with a separate digital corpus with differing orthography. This includes similar excerption-based historical dictionaries that have been digitized, and smaller projects such as glossaries of individual texts or subcorpora. The key foundation to this approach is an accurate process to completely lemmatize the words in the text.

## References

1. Wills, T., Johannsson, E.: Reengineering an Online Historical Dictionary for Readers of Specific Texts. In: Kosem, I. (ed.) Electronic lexicography in the 21st century: Smart lexicography: Proceedings of the eLex 2019 conference, pp. 116–129. Lexical Computing, Brno (2019).
2. Unger, C. R. (ed.): Thomas Saga Erkibyskups: Fortælling om Thomas Becket Erkebiskop af Canterbury: To Bearbeidelser samt Fragmenter af en tredie. Kristiania (1869).
3. Medieval Nordic Text Archive (Menota), https://menota.org, last accessed 2020/12/02.
4. Haugen, O. E.: The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources. Version 3.0. Medieval Nordic Text Archive, Bergen (2019). http://www.menota.org/handbook.xml
5. Menota Public Catalogue, https://clarino.uib.no/menota/catalogue, last accessed 2020/12/02.
6. Wills, T., Johannsson, E., & Battista, S.: Linking Corpus Data to an Excerpt-based Historical Dictionary. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.): Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, pp. 979–987. Ljubljana: Ljubljana University Press, Ljubljana (2018).
7. Lomicka, L.: To gloss or not to gloss: An investigation of reading comprehension online. Language learning & technology 1(2), 41–50 (1998).
8. Abraham, L. B.: Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. Computer Assisted Language Learning 21(3), 199–226 (2008).