The Case for Normalization: Linking Lexicographic Resources for Icelandic

Kristín Bjarnadóttir

The topic of this paper is the linking of two major lexicographic resources on Icelandic, the Dictionary of Old Norse Prose (ONP) in Copenhagen, and the Written Language Archive (WLA) at the Árni Magnússon Institute for Icelandic Studies in Reykjavík. The resources would be linked by normalizing the headwords from the ONP to the standard used for Modern Icelandic spelling by the use of a spellchecker originally designed for use on Modern Icelandic. This is possible, as the major differences in word forms between the two sources are mostly a matter of changes in spelling, whereas the morphology of Icelandic has been relatively unchanged through the centuries. Future work would consist of normalizing the citation collections themselves, thus creating a valuable resource linking variant word forms to their headwords or lemmas.

1. Introduction

The two major, longstanding lexicographic projects on Icelandic are the Dictionary of Old Norse Prose (ONP) in Copenhagen, and the work on a historical dictionary of Icelandic, as manifested in the Written Language Archive (WLA, Ritmálssafn Orðabókar Háskólans), in Reykjavík.[1] The line of demarcation between the projects is 1540, i.e. the year of publication of the first printed book in Icelandic. The projects were started in 1939 and 1944 respectively, and they are independent of each other, both in scope and methods. The topic of this paper is the linking of the data from these two

[1] The Dictionary of Old Norse Prose (ONP), originally established by The Arnamagnæan Commission in Copenhagen, now part of the Department of Nordic Research, University of Copenhagen (Ellert Þór Jóhannsson et al., this issue), and The Written Language Archive, originally at The Institute of Lexicography (Orðabók Háskólans), now part of The Árni Magnússon Institute for Icelandic Studies, Reykjavík. For further information, see the websites: onp.ku.dk & arnastofnun.is. Thanks are due to the staff at the ONP and to Jón Friðrik Daðason.

Nordiske Studier i Leksikografi 13 (2016): 79-88

resources by the normalization of the data in the ONP to standardized Modern Icelandic, as used in the lemmatization in the WLA. The first stage is the linking of the lemmas in both sources described below, but future work would include normalization of the citations themselves by automatic methods, thus creating a combined resource useful both in lexicography and in research on and access to data on the language in general.

2. The sister projects, ONP and WLA

Both projects, the ONP and the WLA, are traditional archives of dictionary slips, collected through decades in the latter part of the 20th century and gradually made available to users online, the ONP as an edited dictionary, the WLA as a collection of citations. The headwords in the ONP are normalized in accordance with the guidelines of the classic manner of Old Icelandic/Old Norse dictionaries, such as Fritzner (1867) and Cleasby & Vigfússon (1874), i.e. to the so-called 'standard Old Norse spelling' ('samræmd stafsetning forn'), traditionally used in editions of medieval Icelandic texts. Figure 1 shows a sample from an entry in the ONP.

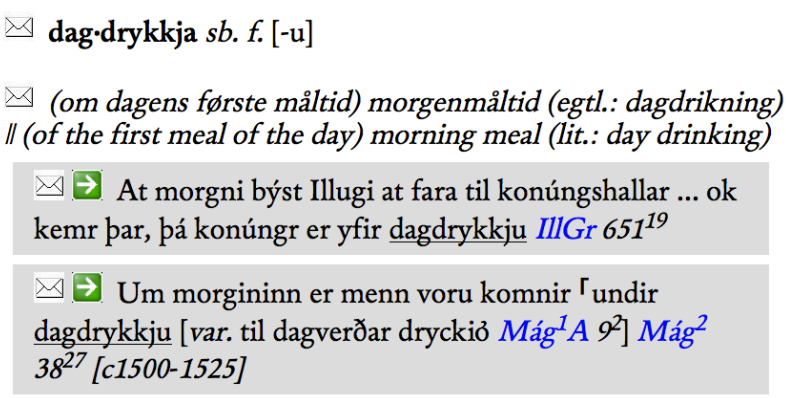

Figure 1: A sample entry from the ONP.

The headwords in the WLA are standardized to Modern Icelandic. For *dagdrykkja* the result is the same for the ONP and the WLA. Examples of differences in spelling are shown in Table 2.

Leitarorð: dagdrykkja

Raða eftir: undanfarandi orði eftirfarandi orði **aldri**

Dæmi 1 - 1 af 1
Síða 1 af 1

Til baka í "dagdrykkja" Ný leit

Nr	Dæmi	Orðmynd	Heimild
1	Þrátt fyrir vanmetakennd og **dagdrykkju** er Guðný ekki búin að vera.	dagdrykkju	TímMM 1980, 254 Aldur: 20s

Dæmi 1 - 1 af 1 Til baka í "dagdrykkja" Ný leit

Figure 2: A sample entry from the WLA.

The citations in both archives are shown with the original spelling from each source, ranging from editions of manuscripts from the 12th century to texts from the period 1980–1985, when traditional excerpting at the Institute of Lexicography was stopped and the focus shifted to the collection of computerized texts. The word forms of the headwords are clearly distinguished in the citations as seen in the examples from the websites in Figures 1 and 2, and links to the sources are given, making both the ONP and the WLA invaluable resources of lemmatized word forms. This data is a valuable language resource in itself, as it can be used to improve NLP tools by linking word forms irrespective of spelling.

3. The vocabulary of the ONP and the WLA

The respective number of headwords is 65.000 for the ONP, and over 600.000 for the WLA. Table 1 shows the proportions of headwords beginning in D in the automatic matching of the ONP and the WLA:

Headwords in WLA, total	16,815
Headwords in ONP, total	1,528
Headwords in ONP & WLA	853
Headwords in ONP only	675
Headwords in WLA only	15,962

Table 1: Figures for headwords beginning in D.

Changes due to discrepancies in lemmatization between the ONP and the WLA are not included in the figures in Table 1. These are due to different practices in the two projects, as in the creation of artificial headwords. This is the case when singular forms of nouns are used as headwords for nouns that only appear in the plural in (e.g., ONP *órr* masc. sg. vs. WLA *órar* masc. pl. 'fantasy, delusion'), and when verbs are lemmatized as infinitives (active voice) in spite of only appearing in the mediopassive and/or as past participles. This method is traditional in older Icelandic dictionaries and it is more common in the ONP than in the WLA where it is rare. Discrepancies in the analysis of word class, such as adjectives vs. past or present participles of verbs, can also result in mismatches between the ONP and the WLA. Differences in word formation can also lead to mismatches, as in *dagaljós* and *dagsljós*, where the first constituent of the compound is either in the genitive singular or plural, without change of meaning. Table 2 shows differences in spelling and word formation in headwords, and one instance of a difference in meaning, i.e., in *dagdrykkja*.

ONP	WLA	Translation (ONP)	Notes
dáðalauss	*dáðalaus*	'amoral'	cf. *dáðlaus*
dáðlauss	*dáðlaus*	'without virtue ...'	cf. *dáðalaus*
dáendi	*dáindi*	'fascination ..., miracle'	
daga·fjǫlði	*dagafjöldi*	'number of days'	
dagaljós	[→dagsljós]	'daylight'	cf. WLA: *dagsljós*
dagsljós	*dagsljós*	'daylight'	
dagdrykkja	*dagdrykkja*	ONP: 'morning meal'	
		WLA: 'drink during the day'	

Table 2: Examples of headwords from ONP and WLA.

By resolving such discrepancies, the number of headwords from the ONP not found in the WLA can be reduced to a degree, but this work has to be done manually and it is quite time-consuming.

As the vocabulary of the ONP is not particularly large, it would be quite feasible to normalize and link the headwords manually, but the process can be made considerably faster by using automatic methods. This process can also serve as a part of the much more challenging project of normalizing running text from all periods of Icelandic language history.

The WLA was from the outset meant to be a historical dictionary for Icelanders and others studying the language from the 16th century onwards, and accordingly the headwords were normalized to Modern Icelandic. The headwords in the ONP are also normalized, but to a different standard of 'normalized Old Icelandic'.

4. The cohesion of Icelandic morphology

The actual point of demarcation between the ONP and the WLA is unwarranted as far as language history is concerned, being set at the date of the first printed book in Icelandic, Oddur Gottskálksson's translation of the New Testament (AD 1540), which is arbitrary from the point of linguistic changes in Icelandic.

The morphology of Icelandic has remained relatively stable through the centuries, both word formation and inflection, but the changes in spelling are extensive enough to make older texts very 'difficult' for modern day Icelanders, without training. The changes in spelling are regular ('predictable') from old to new, but not vice versa. With normalized modern spelling, older word forms are for the most part easily manageable for modern readers. In the first ten lines of the oldest Icelandic manuscript *Reykjaholtsmáldagi* (ca. AD 1150), 22 % of the word forms are in fact identical to the modern word forms, but the figure rises to 78 % by the use of 4 sets of simple rules, such as changing *o/v* to *u* in endings, changing the ending *-r* to *-ur*, and changing *þ* to *ð* except when word-initial (cf. Figure 3).

Original:
Tıl kırkıo lıgr ıravkıaholte heıma land meþ ollo᷑ landſ nẏtıom
þar fẏlgıa kẏr tottogo·grıþungr tuevetr· xxx·a·oc hundraþ·

Modern character set:
Til kirkio ligr iravkiaholte heima land meþ ollom lands nytiom
þar fylgia kýr tottogo griþungr tvevetr xxx a oc hundraþ.

Modern spelling:
Til kirkju liggur í_Reykjaholti <heimaland> með öllum <landsnytjum>.
Þar fylgja kýr tuttugu, griðungur tvævetur, 30 áa og hundrað.

Translation:
To the church in Reykjaholt belongs the home farm with all its produce. Along with that go twenty cows, a two-year old bull, and one hundred and thirty ewes.

Figure 3: The first two lines from Reykjaholtsmáldagi (AD 1250).

The spellchecker Skrambi described below makes this work feasible.[2]

4.1 Changes in spelling

The historical changes in Icelandic spelling are quite extensive, but here a small demonstration of the variants appearing in the entries of the word *drykkja* in the ONP and the WLA must suffice (cf. *dagdrykkja* in Figures 1 and 2 above).

- ONP: *dryckiom, drvckio, drvkkio, dryckiar-, drykkior, dryckio, dryccior, dryckiu, dryckiona, drykkivr, drykkju, ðrykkiu, dryckiv* ...
  - [ck|k|kk|cc] → kk
  - i → j; v → u (in endings)
  - [u|o|v] → u
  - ð → d
- WLA: *dryckia, dryckiu, drykkja, drykkju* ...
  - ck → kk
  - i → j

Figure 4: Spelling variants in *drykkja* in ONP and WLA.

The changes in the ONP are much more extensive than in the WLA, as expected considering the temporal scope, but even the variants in the WLA are sufficient to make (untrained) modern readers shy away. In both cases, however, the modern equivalent can be found by the use of a spellchecker.

[2] The spellchecker will be applied to editions as prepared by philologists who will resolve all problems relating to paleography. An unresolved problem is the different treatment of word boundaries, as seen in Figure 3. The spellchecker can be expected to cope with that to a degree, as word boundaries are also a problem in modern spelling.

5. Automatic spelling normalization

Experiments with automatic spelling normalization with a spellchecker have been ongoing at the Árni Magnússon Institute for Icelandic Studies for some years, mainly working on 19th century texts (Svavarsdóttir et al. 2014, Daðason et al. 2014). The spellchecker Skrambi (Daðason 2012) is based on a noisy channel model. Given an unknown word, it will find all similar word forms in the Database of Modern Icelandic Inflection (Bjarnadóttir 2012). These potential corrections are then ranked according to their frequency (as determined from a large corpus) as well as the probability of the specific character edits (i.e. substitutions, deletions, insertions and transpositions) that were required to transform (misspell) the suggested word into the unknown word. Skrambi is used in tandem with other NLP tools for the analysis of Icelandic texts, as in PoS tagging, lemmatization, and compound splitting, but all currently available tools are made to be used on Modern Icelandic only. As NLP tools are generally expensive to make and often rely on extensive data, the advantage of normalizing older texts in order to be able to analyze them and make them searchable by using existing tools is obvious. The crux of the matter is whether the changes in spelling are regular enough to make this possible. This proved to be the case with the 19th century texts referred to above, even though the spelling of some of them was highly eccentric and idiosyncratic. The next step was to experiment with even older texts.

When Skrambi is extended to normalize older texts, it simply treats the older variants as errors and tries to correct them by comparison with modern variants. Skrambi utilizes machine learning techniques, i.e. it learns from the errors it corrects, and access to quantities of variant forms that are attributed to source and thus dated is of immense value. This is where data from both the ONP and the WLA is of great importance, not only in this project but in future work with Icelandic texts from all ages.

6. The dating of words in the WLA

As the WLA is excerpted in the traditional manner of reading and producing dictionary slips of noteworthy examples of usage, the dating of words can

sometimes be quite faulty. To name an example, the headword *eiginmaður* 'husband' only has 4 citations in the WLA, all with older (non-standardized) spelling, which in itself makes them noteworthy. The modern spelling does not appear at all under the headword, giving insufficient information of a common noun. The citations in the WLA have recently been made accessible for text search, and there are in fact there 124 additional examples of the modern spelling of *eiginmaður* in the WLA, contained in citations under different headwords. The dating of headwords and variants in the WLA should therefore be taken with a grain of salt, as the traditional method of excerption is not a guarantee of correctly dated words.

The general public in Iceland is not sufficiently well aware of the ONP and of the necessity of consulting both the WLA and the ONP when trying to date words. In Iceland, the WLA sometimes has the immense status of being the definite source of information on a word. People will therefore jump to conclusions, and inane as it may sound, the answer to the question "How old is the word *hundur* 'dog'?" may very well be "1540", as the oldest citation of the word in the WLA is indeed from the 1540 New Testament.

Other lacunae in the WLA are more difficult to spot, as seen in Table 3, which shows the examples of headwords from the letter D in the ONP in which the dating in the WLA appears misleading.

Dating	*No. of entries*	*Example*	
No date	7	*dammur* n.masc.	'dam'
15 th C	2	*drykkjarhorn* n.neut.	'drinking horn'
16 th C	280	*dauðasvefn* n.masc.	'sleep of death'
17 th C	195	*danskur* adj.	'Danish'
18 th C	89	*dauðdagi* n.masc.	'day/manner of death'
19 th C	202	*Danaveldi* n.neut.	'Kingdom of Denmark'
20 th C	78	*dagdrykkja* n.fem.	'day drinking'

Table 3: The dating of words from the ONP in the WLA. Examples from D.

As the ONP shows clearly, the adjective *danskur* 'Danish' is not a 17th century word, and *dauðdagi* 'day of death' is not from the 18th century. The word *dagdrykkja* (cf. Figures 1 and 2) is an interesting case where the 20th

century word in the WLA may indeed be a ‘new’ word, as the meaning ‘day drinking’ (as a sign of alcoholism) is different from the meaning in the ONP, where it signifies ‘breakfast’ (Orkneyinga saga). The compound is perfectly regularly made, however, then and now, made with the same constituents.

The users of the WLA would be well served if they could access the ONP directly, without technical hitches. To make that possible, the lists of headwords must be connected, and the easiest way is to normalize both to the same standard, which in the case of the WLA is Modern Icelandic.

7. Conclusion

The aim of the approach described in this paper is to maximize the usefulness of linguistic resources and the NLP tools created in the last few years for the analysis of Modern Icelandic texts. Language resources and tools are expensive to make, and in the case of both the ONP and the WLA, the archives have been decades in the making. Both of them are magnificent resources that should be easily accessible to as many people as possible, in as many ways as possible. If existing NLP tools can be adapted to cope with older forms of Icelandic, the general public would gain better access to the history of individual words, and scholars (i.e., philologists, lexicographers, linguists, historians, etc.) would be able to utilize the texts to a fuller extent. By the anchoring of spelling variants and lemmas, a new resource for research would be created. The benefits of the normalization would be better access to Icelandic vocabulary throughout history, in lexicography, in research on the language in general, and for the general public, as in the creation of search engines coping with Icelandic of all ages.

Bibliography

Dictionaries

Cleasby, Richard & Guðbrandur Magnússon (1874): *An Icelandic-English Dictionary*. Oxford: Clarendon Press.

Fritzner, Johan (1867): *Ordbog over Det gamle norske Sprog*. Kristiania: Feilberg & Landmark.

ONP = *Dictionary of Old Norse Prose* [*Ordbog over det norrøne prosasprog*]: <onp.hum.ku.dk>.
WLA = The Written Language Archive: <arnastofnun.is/page/gagnasofn_ritmalssafn>.

Other references

Bjarnadóttir, Kristín (2012): The Database of Modern Icelandic Inflection. In: *LREC 2012 Proceedings: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 – AfLaT 2012*, 13-18.

Daðason, Jón Friðrik (2012): *Post-Correction of Icelandic OCR Text*. <hdl.handle.net/1946/12085>.

Daðason, Jón Friðrik, Kristín Bjarnadóttir & Kristján Rúnarsson (2014): The Journal *Fjölnir* for Everyone: The Post-Processing of Historical OCR Texts. In: *Proceedings of Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage*, 55-62. LRT7HDA. LREC 2014, Reykjavík.

Gottskálksson, Oddur (1540): *Nýja testamentið*. Roskilde.

Gunnlaugsson, Guðvarður Már (ed.) (2000): *Reykjaholtsmáldagi*. Reykholt, Reykholtskirkja, Snorrastofa.

Jóhannsson, Ellert Þór & Simonetta Battista (2016): Ordbog over det norrøne prosasprog online – struktur og brug. This publication, xx-xx.

Svavarsdóttir, Ásta, Sigrún Helgadóttir & Guðrún Kvaran (2014): Language Resources for Early Modern Icelandic. In: *Proceedings of Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage*, 19-25. LRT7HDA. LREC 2014, Reykjavík.

Kristín Bjarnadóttir
reasearch lecturer
The Árni Magnússon Institute for Icelandic Studies/Háskóla Íslands
Laugavegi 13
IS-101 Reykjavík
kristinb@hi.is