Kristín Bjarnadóttir
The Árni Magnússon Institute for Icelandic Studies
University of Iceland

The Case for Normalization:
Linking Lexicographic Resources for Icelandic

13. Konference om Leksikografi i Norden
København
20.5. 2015
Before or after 1540?

“How old is the word hundur?” (‘dog’)

“I searched in the “Dictionary” (i.e. Orðabók Háskólans). It’s from 1540.”

Hypothetical question, but it holds true for rarer words
Topic: Access to Information

The Written Language Archive (WLA) and The Dictionary of Old Norse Prose (ONP). Line of demarcation: 1540

The lists of entries: Proportions, lemmatization, dating of words

The normalization: The cohesion of Icelandic through the centuries, modern standardized spelling, the spellchecker

Conclusion: The benefits of the normalization are better access to Icelandic vocabulary, both in the ONP and WLA, and in research on the language in general.
Two Sister Projects

The Dictionary of Old Norse Prose, University of Copenhagen
- Period: up to 1540
- Number of headwords: approx. 65,000
- Established in 1939, by The Arnamagnæan Commission
- Purpose: Replacing the two major 19th C dictionaries, Fritzner and Cleasby & Vigfússon
- The ONP is now a part of The Department of Nordic Research (Nordisk Forskningsinstitut)

The Written Language Archive, The Institute of Lexicography, Reykjavík
- Period: 1540 to the 1980s
- Number of entries: over 600,000, 2.5 million citations
- Established in 1944, as Orðabók Háskólan (The Institute of Lexicography)
- Now part of The Árni Magnússon Institute for Icelandic Studies

Examples:
**dag-drykkja** *sb. f. [-u]*

*(om dags første måltid) morgenmåltid (egtl.: dagdrikning) // (of the first meal of the day) morning meal (lit.: day drinking)*

- At morgni býst Illugi at fara til konúngshallar ... ok kemr þar, þá konúngr er yfir dagdrykkju **IllGr 651**

- Um morganinn er menn voru komnir «undir dagdrykkju [var. til dagverðar dryckið **Mág**¹ A 92] Mág² 38²² [c1500-1525]*

- Iarl hafði stádit upp snimma, ok höfdu þeir ... farit sudr a eyna at veida otr ... þeir ætludu heim til dagdrykkju **Orkn 187**¹⁸ [c1387-1395]

- Nú líðr af náttin ok kemr dagverðardrykkja. Ok eptir dagdrykkju lætr konungr kalla til sin sendiboðana **Rém**² 335²

**Lítt.** *Ejder 1967 [KLNM 12] 120.*

**Gloss:** *EJ; CIV; Fr; CIVSuppl; NO; (Bl)*
ONP Citation: *dagdrykkjja*

‘*morning meal*’ (lit. ‘*day drinking*’)

---

Iarl hafdi stadit upp snimma, ok hofdu þeir ... farit sudr a eyna at veida otr ... þeir ætludu heim til dagdryckiu *Orkn 187*¹⁸ [c1387-1395]
Ritmálssafn
Orðabókar Háskólanys

Niðurstoða leitar

dagdrykkja
nafnorð kvenkyn

Dæmi í ritmálssafni frá 20s
Heimild elsta dæmis: TimMM 1980, 254
Fjöldi tildækra dæma: 1

Athugið: Upplysingar um fjölda tildækra dæma miðast við þau dæmi sem komin eru í gagnagrunnin. Í ritmálssafni kunna að vera fleiri dæmi.

Sjá dæmi  Sjá orðmyndir í dæumum
Ný leit

Tilvitan í efni Ritmálssafns
WLA Citation: dagdrykkja
“In spite of a feeling of inferiority and day drinking ... “
Caveat:
Beware of semantic changes

*dagdrykkja*:

Modern Icelandic: ... a sign of alcoholism
Orkneyinga saga: breakfast!
The headwords in ONP are standardized in the classic manner of Old Icelandic.
The headwords in WLA are standardized to Modern Icelandic.
Combining the two lists is relatively simple.

The actual difference between the sources is not quite as great as shown in the figures here, due to discrepancies in lemmatization, etc.

<table>
<thead>
<tr>
<th>Headwords in ONP:</th>
<th>1,528*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headwords in both ONP and WLA:</td>
<td><strong>853</strong></td>
</tr>
<tr>
<td>Headwords only in ONP:</td>
<td>675</td>
</tr>
<tr>
<td>Headwords only in WLA:</td>
<td>15,962</td>
</tr>
<tr>
<td>Headwords in WLA:</td>
<td>16,815</td>
</tr>
</tbody>
</table>

\[
\text{WLA total: 600,000} \quad \text{ONP total: 65,000}
\]
Discrepancies in Lemmatization

- Singular vs. plural in nouns:
  - ONP draumórr sg. vs. WLA draumórar pl. ‘dream fantasies, delerium’. The singular is a lexicographer’s generalisation.

- Active voice vs. mediopassive and/or past participle in verbs:
  - ONP daufheyra vs. WLA daufheyrast, daufheyrður ‘ignore, pretend not to hear’. The active voice is a generalisation.
  - ONP ótta vs. WLA óttast ‘be afraid of’. The active voice is attested in Old Icelandic but not in the modern language.

- Checking the discrepancies must be done manually.
# The Dating of Words from ONP in WLA

<table>
<thead>
<tr>
<th>Dating</th>
<th>Citations</th>
<th>Example</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>No date</td>
<td>7</td>
<td>dammur n masc</td>
<td>‘dam’</td>
</tr>
<tr>
<td>15th C</td>
<td>2</td>
<td>drykkjuhorn n neut</td>
<td>‘drinking horn’</td>
</tr>
<tr>
<td>16th C</td>
<td>280</td>
<td>dauðasvefn n masc</td>
<td>‘sleep of death’</td>
</tr>
<tr>
<td>17th C</td>
<td>195</td>
<td>danskur adj</td>
<td>‘Danish’</td>
</tr>
<tr>
<td>18th C</td>
<td>89</td>
<td>dauðdagi n masc</td>
<td>‘day of death’ *</td>
</tr>
<tr>
<td>19th C</td>
<td>202</td>
<td>Danaveldi n neut</td>
<td>‘Kingdom of Denmark’</td>
</tr>
<tr>
<td>20th C</td>
<td>78</td>
<td>dagdrykkja n fem</td>
<td>‘day drinking’</td>
</tr>
</tbody>
</table>

**danskur** ‘Danish’ is not a 17th Century word in Icelandic!

**hundur** ‘dog’ was not a new word in 1540. 
Linking the ONP and WLA makes sense.
• Relatively little changes in morphology, both in word formation and inflection.
• Sufficient changes in spelling to make the older texts very ‘difficult’ for Icelanders, without training.
• The changes in spelling are regular (‘predictable’) from old to new (but not vice versa).
• Modern Icelandic spelling is standardized, from the early 20th C only. Normalization is therefore also needed for the WLA.
• The project described here consists of normalization of word forms already rendered in a modern character set.
Example of Icelandic Spelling:
Reykjaholtsmáldagi AD 1150, Lbs 340 b 4to

Til kirkio ligr í ræk ciąholte heima land meþ ollo\lánds nýtjom
þar fylgia kyr tottogo griþungr tvevetr\xxx a oc hundrap.

Original:
Til kirkio ligr í ræk ciąholte heima land meþ ollo\lánds nýtjom
þar fylgia kyr tottogo griþungr tvevetr\xxx a oc hundrap.

Modern character set:
Til kirkio ligr í rækciaholte heima land meþ olom lands nytjom
þar fylgia kyr tottogo griþungr tvevetr xxx a oc hundrap.

Modern spelling:
Til kirkju liggur í_Reykjaholti <heimaland> með öllum <landsnytjum>.
þar fylgja kyr tuttugu, griðungur tvævetur, 30 áa og hundrað.
The Spelling of (dag)drykkja

• **ONP**: dryckiom, drvckio, drvkkio, dryckiar-, drykkior, dryckio, dryccior, dryckiu, dryckiona, drykkivr, drykkju, drykkiu, ðryckiiu, dryckiv ...

• **WLA**: dryckia, dryckiui, dryckju, drykkja, drykkju ...

• **ONP**:
  - \([ck|k|kk|cc] > kk\)
  - \(i > j\), \(v > u\)
  - \([u|o|v]$* > u\)
  - \(v > y\)
  - \(^\dd > d\)

• **WLA**:
  - \(ck > kk\)
  - \(i > j\)

• Modern Icelandic: **drykkja**  
  cf. BÍN
The Database of Modern Icelandic Inflection

Beygingarlysing íslensks nútímmamáls
Stofnun Árna Magnússonar í íslenskum fræðum

Rístjóri Kristín Bjarnadóttir

<table>
<thead>
<tr>
<th>Eintala</th>
<th>Fleirtala</th>
</tr>
</thead>
<tbody>
<tr>
<td>án greinis</td>
<td>án greinis</td>
</tr>
<tr>
<td>með greini</td>
<td>með greini</td>
</tr>
<tr>
<td>Nf.</td>
<td>drykkja</td>
</tr>
<tr>
<td>Pf.</td>
<td>drykkju</td>
</tr>
<tr>
<td>Bgf.</td>
<td>drykkju</td>
</tr>
<tr>
<td>Ef.</td>
<td>drykkju</td>
</tr>
</tbody>
</table>
The normalization is done automatically, treating the spelling variants as errors. The spellchecker is called **Skrambi**

- Skrambi is based on a noisy channel model
- Skrambi uses data from the Database of Modern Icelandic Inflection (BÍN)
- Skrambi is used as
  - an ordinary spellchecker for Modern Icelandic
  - to correct Optical Character Recognition errors
  - for normalization of historical texts
- Skrambi is programmed by Jón Friðrik Daðason
The WLA citations are a meticulously dated resource with original spelling, providing extremely valuable samples of text, not just of the respective headwords. Same applies to the ONP.

The WLA was excerpted in the traditional manner, by reading.

Example from the WLA: eiginmaður ‘husband’

- 4 citations, all of older spelling
- Searching the citations as text yields 124 additional examples with modern spelling.
- With normalization the number increases, as older forms are added.
<table>
<thead>
<tr>
<th>Nr</th>
<th>Dæmi</th>
<th>Orðmynd</th>
<th>Heimild</th>
</tr>
</thead>
</table>
| 1  | ef konan forlætr sinn **eigenn mann**.    | **eigenn mann** | OGottNýjat  
Aldur: 16m  
Mk 10, 12 |
| 2  | byscupi byriar at vera [ [...] ] einnrar kuinnu **eiginn manni**. | **eiginn manni** | OGottNýjat  
Aldur: 16m  
1Tm 3,  
1Tm 3, 2 |
| 3  | byscupi byriar at vera [ [...] ] einnrar kuinnu **eiginn manni**. | **eiginn manni** | OGottNýjat  
Aldur: 16m  
Mk 10, 12 |
| 4  | ef konan forlætr sinn **eigenn mann**.    | **eigenn mann** | OGottNýjat  
Aldur: 16m  
Mk 10, 12 |

Takmarka listann við dæmi sem innihalda:  
Leita
Searching the WLA Citations as Text
*eiginmaðr sb. m. (CIV "K. Á. 136" => eigin mann ÁKr 39²³; tolkes som to ord) \( \rightarrow \) eigin adj.

**Gloss:** EJ; CIV; Fr; LP; NO; (Bl)
The Benefits of Linking ONP and WLA, and of Normalization

- Access to Icelandic from all periods of its history. Both ONP and WLA are magnificent resources.
- The general public: Better idea of the history of individual words.
- Scholars: Access to research material, by the anchoring of spelling variants to modern word forms and lemmas.
- Modern NLP tools generally assume modern spelling.
- Normalization greatly enhances the scope of search engines.
- ...
Thank you for your attention

And thank you, Þorbjörg, Aldís, Ellert, Simonetta, ...
at the ONP

Kristín Bjarnadóttir
kristinb@hi.is

May 20 2015
NFL Copenhagen


• Rms
• ONP
• Menota
• .
• .
• .
Normalization of the ONP Citations:
dagdrykkja... Exhaustive excerption?

Click Sigurður Nordal 1913-1916 [STUAGNL 40] for the full bibliographical entry and copyright.

Main Text
p. 186 (separate window)  p. 187 (separate window)  p. 188 (separate window)

Iarl hafði státt upp snírama, ok hofdu þeir ... farit sudr a eyna at veida otr ... þeir ætludu heim til dagdryckiu Orkn 187¹⁸ [c1387-1395]

otr, er læ i urdninn undir hofdanum; þeir ætludu heim til dagdryckiu. Byrdingsmenn reru at landi, ok spurduz þeir almætra² tidenda. Sogdu³ byrdingsmenn, hvatan þeir væri at komnir; fretta þeir ok, ²⁰ hvar iarl væri. Þeir segia, at hann væri þar a