Digitizing the Icelandic-Danish Blöndal Dictionary

Steinþór Steingrímsson[1]

[1] The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland
steinst@hi.is

Abstract. The Icelandic-Danish dictionary, compiled by Sigfús Blöndal in the early 20th century is being digitized. It is the largest dictionary ever published in Icelandic, containing in total more than 150,000 entries. The digitization work started with a pilot project in 2016 resulting in a comprehensive plan on how to carry out the task. The paper describes the ongoing work, methods and tools applied as well as the aim of the project and rationale. We opted for using OCR and not double-keying, which has become common for similar projects. The entries are annotated with XML-entities, using a workbench built for the project. We apply automatic annotation for the most consistent entities, but other annotation is carried out manually. The data is then exported into a relational database, proofread and finally published. Publication date is set for spring 2020.

Keywords: Digitization, Dictionaries, Icelandic.

1 Introduction

This paper describes on-going work in digitizing the Icelandic-Danish dictionary by Sigfús Blöndal, published in 1920-1924, and a supplement published in 1963.

The Icelandic-Danish dictionary by Sigfús Blöndal, published in 1920-1924, was a very large dictionary project at the time, taking more than 20 years to finish (Blöndal, 1920-1924). The dictionary's 1100 pages, 21 by 28 cm contained 115 thousand entries, printed in small type in two columns (Jónsson, 1997). Furthermore, a supplement was published in 1963, containing more than 39 thousand entries. It is still the most extensive published Icelandic dictionary with more than 150,000 total entries. It has been reprinted two times, last time in 1980.

The dictionary was created using state-of-the-art methods of the time, paper cards created for the entries, arranged in order to be typeset by hand and printed. For many decades now, most dictionaries are prepared on computers allowing for easier search and research using all the different fields that constitute the entries.

Digitizing old paper dictionaries, scanning them and converting into a format that gives useru the same control they have when using dictionaries compiled with the most modern methods, sometimes called retro-digitization, is not as simple as it may seem (cf. Maxwell and Bills 2017). The first step is either to key the text in by hand or using Optical Character Recognition (OCR). Next, the dictionary's visual layout has to be converted into lexicographically structured computer readable format so it can be used as a modern digital dictionary. This introduces possibilities to sort, organize and filter the dictionary entries using different constituents, i.e. dialects, usage,

parts of speech etc. Finally the new structured format has to be prepared for use, both for research and general lookup and search.
The digitization of the Blöndal dictionary started in 2016 when a pilot project was initiated, investigating how the process could best be optimized. The dictionary was scanned in high resolution and we experimented with OCR. We found that after extensively training the OCR we achieved satisfactory results. Most of the OCR errors were in certain fields, the ones printed in bold or italics in the dictionary. Automatic methods to flag likely errors in these fields will be applied and some post-processing will be required. We created a simple workbench for defining XML elements and annotating the data using these elements. After annotation the entries are automatically imported into a relational database. During the pilot stage, 15% of the dictionary was annotated. We learned how best to adapt our workbench to the project, so the lion's share of the work would go smoothly, be as fast as possible and yield accurate and correct annotation. The second part of the digitization started in November 2017 and is expected to finish in 2019.
The paper is organized in the following way. In Section 2 the aim of the project is described and the rationale for this work. Section 3 describes the workflow, the OCR process and the tools created for the project for semi-automatic and manual annotation. In Section 4 a brief overview is given of the XML entities defined for the different dictionary entry constituents. Section 5 concludes with discussion about the timeframe of the project.

2 Rationale and aim of the project

Old dictionaries like the Blöndal dictionary can be expected to be out of date to a lesser or greater extent. When planning to retro-digitize the dictionary the first question that came up had to be, why? To that question there is more than one answer. Generally speaking, it still contains a lot of potentially useful information, even for regular dictionary users. The dictionary is after all the largest dictionary compiled for Icelandic. But even more interestingly, by digitizing it we will open up powerful new ways of research. Dialect lexica, citations to real-world examples, usage labelling etc. can all be researched thoroughly, which is an unsurmountable task while the dictionary is only available in its 1100 pages paper format.
Electronic dictionaries are used in a wide field of new research areas such as the growing community of digital humanities and there is a growing demand for digitizing research tools like dictionaries in the context of digitizing the cultural heritage (Schneiker et al. 2009). The results of this digitization project will hopefully also appeal to researchers in these fields.
The aim of the project is to publish a digital version of the Blöndal dictionary online, searchable in the variety of ways modern digital dictionaries allow for. Thus we are concerned about minimizing typographical errors while converting it into a lexicographically structured computer-readable format. The users should be able to use and research the data as they best see fit. In order to accommodate this we created a workflow described in the next section.

3 Workflow

The first step in a digitization project such as this is to get the text into the computer, structured or not. Different projects have adapted one of two methods. Double-keying the text (Haaf et al. 2013), which is commonly carried out in China. This was for example the case with the Campe dictionary in the German Textgrid project (Schneiker et al. 2009). The other method is using OCR. As described in Section 1 we achieved acceptable results in the pilot project, but some post-processing is required to check for errors in fields flagged if it likely contains an error. Proofreading will also be necessary when everything has been annotated and flagged fields have been checked.

After OCR has been applied the text is imported into a workbench tool specially developed for this project. It is a web based tool, allowing users to work on the operating system they are most comfortable with. The tool allows the user to define new entities if necessary, and to tag dictionary entries using defined entities. The entities are arranged in a schema, to make error checking possible. When importing into the workbench the entries are automatically tagged. The automatic tagging only applies to selected entities, the ones most consistent in the dictionary. The entries are then saved in the tools database for editing. In the editing process the user checks if the automatic tagging is correct, fixes what has to be fixed and tags other entities manually. The user can also fix OCR errors if they are spotted in the process. When each entry is saved the workbench tool makes sure it conforms to the schema and highlights errors if it finds any, for the user to fix.

The correctly tagged entries are then exported to a relational database, from which they can be accessed in a web browser, for viewing, browsing or searching. While the project is ongoing the website is only open to those working on the project, for testing and evaluating the design and user interface of the on-line dictionary. When the project is finished a website will be opened, giving everyone access to the data.

4 Structuring the entries

The entry structure can be quite complicated. It is therefore important to define the right entities for tagging and to be able to automatically check if the tagging complies with the schema. Figures 1 and 2 show a simple entry in the dictionary and how it is tagged.

```
<f>
  <fhl-flettu>afspyrnu</fhl-flettu>
  <shl-flettu>forátta</shl-flettu>
  <hljodritun>[]</hljodritun>
  <formdeild>f.</formdeild>
  <donsk-thyding>forrygende Uvejr.</donsk-thyding>
  <shl-flettu>-rok</shl-flettu>
  <hljodritun>[]</hljodritun>
  <formdeild>n.</formdeild>
  <shl-flettu>-veður</shl-flettu>
  <hljodritun>[]</hljodritun>
  <formdeild>n.</formdeild>
  <donsk-thyding>voldsom, rasende Storm.</donsk-thyding>
</f>
```

Fig. 1. An example of how a simple entry is tagged.

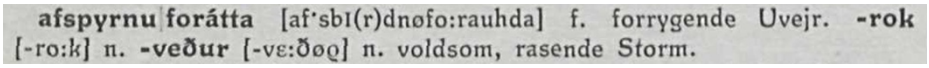
afspyrnu forátta [af'sbɪ(r)dnøfo:rauhda] f. forrygende Uvejr. -rok [-ro:k] n. -veður [-vɛ:ðøɋ] n. voldsom, rasende Storm.

Fig. 2. A simple entry in the dictionary.

This entry only uses five different entities, but in the pilot project 33 different entities were defined. Many of these are rare, but some are quite common. When the entities are exported to the database they are stored in a structure that allows for easy rebuilding of the original entry, and for powerful organization of the data according to each user's needs. For this project Icelandic is used for naming all the entities. These names can easily be translated or redefined for other languages. The workbench tool will likely be tweaked somewhat during the project, in order to make it more efficient. By the end of the project it will be assessed whether it can easily be repurposed for similar projects, and if so, released with an open license.

5 Timeframe and publication

We timed the effort in the pilot project and estimate the whole process to take 26-30 working months. 1-3 persons will be working on the project at any given time and it is estimated to be finished by fall 2019. The dictionary will be published online. A launch date has already been set, April 23 2020, on the one hundredth anniversary of publication of the dictionary. The website will contain papers and various writings by scholars about the dictionary from the last 100 years, as well as an interface to the dictionary that allows searching, browsing or filtering for all the different entry constituents, or fields, giving rise to vast new possibilities in researching the dictionary's content.

References

1. Blöndal, S.: Íslensk-dönsk orðabók. Reykjavík (1920−1924).
2. Haaf, S., Wiegand, F. and Geyken, A.: Measuring the correctness of double- keying: Error classification and quality control in a large corpus of TEI-annotated historical text. In: Journal of the Text Encoding Initiative (2013:4).
3. Jónsson, B.: Stærð orðaforðans í Orðabók Blöndals. In: Orð og tunga 3, pp. 15–19. Orðabók Háskólans, Reykjavík (1997).
4. Maxwell, M., Bills, A.: Endangered Data for Endangered Languages: Digitizing Print dictionaries. In: Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pp. 85–91. Association for Computational Linguistics, Honolulu (2017).
5. Schneiker, C., Seipel, D. and Wegstein, W.: Schema and Variation: Digitizing Printed Dictionaries. In: ACL-IJCNLP '09 Proceedings of the Third Linguistic Annotation Workshop, pp. 82–89. Association for Computational Linguistics, Singapore (2009).