Retro-digitizing Blöndal - Lessons Learnt

Steinþór Steingrímsson

The Árni Magnússon Institute for Icelandic Studies

Íslensk-dönsk orðabók, The Icelandic-Danish Dictionary, compiled by Sigfús Blöndal in the early 20th century has been digitized. The dictionary is the largest dictionary ever published in Icelandic, containing over 150 thousand entries. The digitization work took over four years, it started in 2016 and was published on the web in 2021.

The aim of the project was to publish a digital version of the dictionary, searchable and researchable using all available entry fields, which are abundant in this dictionary, with all from traditional information like word class, definitions, equivalent words and examples to information on regional dialects and a source for the words or examples. In order to do that we had to make multiple methodological choices, sometimes with limited information on the efficacy of each choice. We chose to use OCR for retrieving the text of the entries in digital format. We then labeled the entry parts with XML tags in a custom made tool, automatically if possible but otherwise manually. We imported the entries into a relational database where the text in different fields of every entry were fixed, automatically when possible but often a manual check was needed. Finally we regenerated the entry to be published in html-format, with the option to search or filter by individual fields.

During this process we found that some of our approaches were not working as well as intended, and while some could be altered on our way, others could not. In this paper we describe our methods, what worked well, what could be improved upon, why it needs improvement and how it could be fixed.

Keywords: Retro-digitization; Digitization; Icelandic; Dictionaries