Language resources for early Modern Icelandic

Ásta Svavarsdóttir, Sigrún Helgadóttir, Guðrún Kvaran
The Árni Magnússon Institute for Icelandic Studies / University of Iceland
Reykjavík, Iceland
asta@hi.is,sigruhel@hi.is,gkvaran@hi.is

Abstract

This paper describes the compilation of a language corpus of early Modern Icelandic, intended for research in linguistics and lexicography. The texts are extracted from a digital library, accessible on the website *Tímarit.is*, containing scanned images of individual pages and OCR read text from all Icelandic newspapers and periodicals from that period. In its present form this resource does not fulfill all needs of linguists and lexicographers, due mainly to errors in the digitized texts, the lack of annotation, and limited search possibilities. To create a new language corpus from this text material with the time and money available, methods and tools for automatic or semi-automatic correction of OCR errors had to be developed. The text was to be corrected according to the originals, without any standardization, which poses various challenges in the construction of the corpus. These connect to the correction process itself, the possibilities of using available tools for tagging and lemmatizing, as well as the design of search functions and interface. The solution was to build a parallel corpus with two layers, one with diplomatic text and the other with a standardized modern version of the same text, with mapping between the two.

Keywords: historical corpora, automatic correction, standardization and mapping

1. Introduction

There is great demand for historical language resources for research purposes in lexicography and linguistics. In Iceland, there exist so far only two historical corpora designed primarily with such needs in mind, 'The Icelandic Parsed Historical Corpus' (IcePaHC; (Rögnvaldsson et al., 2012)), spanning the entire history of the Icelandic language from the 12th century onwards, and a corpus of Old Icelandic, i.e. medieval, narrative texts, mainly from the sagas (Rögnvaldsson and Helgadóttir, 2011).[1] The texts in both corpora have a standardized, modern spelling, which has made it possible to handle them with available language technological tools, and facilitates search. At the same time this feature limits the use of these resources. They have proved to be valuable for syntactic, and to a certain extent morphological, research, and they are also useful in lexicography within the limits set by their relatively small size. On the other hand, they can not be relied on in historical investigations of e.g. orthography or phonology.

This paper presents an ongoing project which aims at the building of a corpus of early Modern Icelandic, i.e. 19th and early 20th century language, with non-fictional prose. The texts are extracted from a digital library which includes all Icelandic newspapers and periodicals published in that period (Hrafnkelsson and Sævarsson, 2014). This resource, accessible on the website *Tímarit.is*, contains scanned images of individual pages and a raw version of an OCR read text, i.e. without corrections or adaptation of any kind. The existence of this digitized text material, however defective, together with the immediate needs of several research projects for resources of 19th century language, were the main motivations for the attempt to build the corpus. It was clear that to create a large enough corpus to serve the various needs of the research projects involved, methods and tools for automatic or semi-automatic correction of OCR errors in the texts had to be developed. Some of these projects, aimed at a broad investigation of 19th century language and language use, requested that the text would be corrected according to the spelling of the originals, without standardization of any kind. This poses various challenges in the preparation and construction of the corpus, with respect to the correction precess itself, the possibilities of using available tools for tagging and lemmatizing, as well as the design of the search functions and interface. The solution was to aim at a corpus with two layers, one with diplomatic text and the other with a standardized modern version of the same text, with mapping between the two. The result would be a parallel corpus, where the layers are not two different languages, but two stages of the same language. This would enable the application of language technological tools designed for the modern language, and allow the users to apply the well kown and standardized modern word forms to search the corpus, and get all variants of these word forms as part of the results. The construction of the corpus, due to be completed in 2014 or early 2015, is described in the article.

The organization of the paper is as follows: In chapter 2., we describe the main characteristics of the digital library, which supplies us with the OCR read text, and the reasons why this resource does not fulfil the needs of linguists and lexicographers. Chapter 3., the main part of the paper, deals with the corpus building process. Here we first discuss the general objectives and main requirements for the content and functions of the corpus, then we describe the selection of texts and their extraction from the archives, and after that we explain the correction procedure and the development of methods and tools for correcting the OCR read texts. The last section recounts the present state of the project and the remaining tasks. Finally, there is a short chapter where we draw the main conclusions from the experience of the procedure we have followed in the project.

[1] The corpus is accessible for search at the website `http://mim.arnastofnun.is/index.php?corpus=for`.

2. A digital library of Icelandic newspapers and periodicals

2.1. *Tímarit.is*: Description and objectives

The National and University Library of Iceland has compiled a digital library of Icelandic newspapers and periodicals, and made it available at the website *Tímarit.is*.[2] The collection covers the period from the late 18th century to the present, with a (nearly) complete coverage of Icelandic newspapers and periodicals published before 1920, and a great and increasing selection of titles from then on. No authorization is needed for the pre-1920 material, but later texts are added with the agreement of their publishers. The database currently contains a total of 866 titles, and pages available online are approximately 4.5 million.[3]

The digital library consists of scanned images of each page (pdf-files), with an OCR read text of the respective pages (txt-files) attached to them.[4] The tool applied for the OCR reading is *AbbyyFineReader*.[5] No corrections are made to the OCR read texts, and there are numerous errors in the text files, although their number varies considerably depending on the quality of the original. The metadata documented for each title in the collection includes the following: publication type (journal, newspaper, etc.), language, number of volumes, number of issues, publication period, location(s), publisher(s), (keyword(s)), description, etc.

All the material is contained in a searchable database, and displayed at *Tímarit.is*. Text search is limited to strings of letters (e.g. word forms), one or more at a time. Metadata can to some extent be used to delimit the search, i.e. to a particular title or a certain period. Results are presented Google-style, and the user can reorder them chronologically, or filter them by title and/or period. The user can also choose whether the results return short (<2 lines) or long (approx. 4 lines) snippets of text. The interface only allows the user to view one page at a time, whether it is selected by browsing (by title, year and issue) or as a result of a text search. This applies both to the images and the text attached to them.

The main objective of the *Tímarit.is* database is to make the newspapers and periodicals easily accessible to the research community as well as the general public.[6] It is especially useful as a research tool in many fields of social sciences and the humanities, including history, literature and language studies of various kinds.

2.2. *Tímarit.is* as a resource for language research

The *Tímarit.is* website has been successfully applied as a resource in linguistics and lexicography. It has, in particular, served as a valuable source of examples and citations. As an effective and reliable tool in language research this resource has, however, various limitations and shortcomings in its present form. Due to OCR errors in the texts, examples of words and structures can, for example, be easily missed even if they occur in the material. For the same reason, as well as the lack of annotation, it cannot be trusted that the search returns all and only relevant examples. The database can therefore not be applied in any kind of quantitative research, even if the results may give a vague indication of the existence, (in)frequency, or distribution of a particular word, word form or word combination. Furthermore, the form in which the search results are presented makes it difficult to get an overview of the results. Working with the database is very cumbersome as each page has to be retrieved separately to check the example, and, if it is relevant, it must be copied into another file for analysis. So even if the *Tímarit.is* gives access to much valuable language material, there is a lot to be wished for concerning the form and presentations of the data with respect to linguistic and lexicographic research.

3. A corpus of early Icelandic newspapers and periodicals for the purposes of language research

3.1. Objectives and requirements

On account of the shortcomings of *Tímarit.is* as a resource for language research, as described in 2.2., a separate corpus of early Icelandic newspapers and periodicals is under construction at The Árni Magnússon Institute for Icelandic Studies. This can be seen as a sub-corpus of the *Tímarit.is* archives in the sense that all the digitized text material is extracted from that. The main objectives of the project are to compile a corpus of early Modern Icelandic non-fictional texts, and construct a database and search interface for the corpus that serve the special needs of research projects in linguistics and lexicography, as well as practical tasks such as dictionary making.

The corpus will cover the 19th and early 20th centuries. Texts from that period, especially the first part of it, pose special problems that have to be solved in the process, as they are not standardized, neither with respect to orthography nor morphology, and the spelling can vary greatly from one title to another, or even within the same paper, e.g. from one time to another. One of the requirements for the corpus is that the texts should be presented with their original spelling and word forms, and in the development of tools for a semi-automatic correction of the OCR read text this has to be taken into account. Another requirement is that search in the corpus should be efficient, flexible, and preferably include possibilities that go beyond simple text search for particular words or word forms. The prerequisite for this is that the texts can be grammatically tagged

[2] The website also contains Faroese and Greenlandic newspapers and periodicals, and the compilation of the corpus and the construction of the database is a collaboration between the National and University Library of Iceland, The National Library of the Faroe Islands, and The National and Public Library of Greenland. In this article, however, we are only concerned with the Icelandic material.

[3] Cf. `http://timarit.is/about_init.jsp?navsel=3&lang=en`. The figures include Faroese and Greenlandic texts, but those are only a small minority of the collection (approx. 30 titles). Some of the Icelandic papers, esp. the older ones, were published in Denmark or the Icelandic settlements in Canada.

[4] Cf. `http://timarit.is/view_page_init.jsp?pubId=315&lang=en` for an example.

[5] Cf. `http://finereader.abbyy.com/`.

[6] Cf. `http://timarit.is/about_init.jsp?lang=en`.

and lemmatized, both in order to overcome the problems posed by the many spelling variants of words and word forms, and to make it possible to search not only for word forms but also for grammatical features. Language technological tools for tagging and lemmatizing Icelandic texts are available, but they have been developed for the modern language (Loftsson, 2008), and it is unclear how useful they would be for earlier language stages. They have, however, been applied successfully for the annotation of Old Icelandic texts (Rögnvaldsson and Helgadóttir, 2011), but these texts had already been standardized for publication according to Modern Icelandic standard orthography. In the construction of the present corpus, the plan is to map the corrected texts automatically to a standardized version, i.e. Modern Icelandic standard spelling, to enable the application of the available tools, and the corpus will thus consist of two layers of text, a diplomatic version and a standardized version.

The immediate needs of ongoing lexicographic and linguistic projects, i.e. studies of lexical borrowings in the 19th and early 20th centuries, [7] and an investigation of language variation and language standardization in 19th century Icelandic, [8] have influenced the content and structure of the corpus. Nevertheless the corpus is intended as a general resource for language research, and is not limited to these particular projects.

3.2. The selection of texts and scope of the corpus

The main criterion for the selection of texts for the corpus was that the collection would be sufficiently representative of the genre and the period in question, within the limits of the *Tímarit.is* archives. The genre is, as previously described, newspapers and periodicals. Such texts were, in fact, a substantial part of published Icelandic texts at the time, esp. in the 19th century, and they are a good representative for non-fictional texts in general, covering a variety of topics and sub-genres, such as narratives, news, discussions, advertisements, etc. (and some of them even include some fiction as well). The period that the corpus covers is the 19th and early 20th centuries, until around 1920. The *Tímarit.is* archives, which is the material available to choose from, contain approx. 35,000 issues of 300 Icelandic titles published in Iceland and Denmark during that period.[9] The distribution of published texts over time is, however, far from even, both with respect to the number of titles and the number and size of issues relating to these titles, and there is much less material to choose from in the first part of the 19th century, than there is later. This is in many ways reflected in the corpus as the selection of texts had to be denser for the early 19th century in order to cover the entire period, and it was difficult to ensure the desirable diversity in the older texts. The proportion of extracted texts in each decade over the period, estimated by title and issue, is displayed in Figure 1.

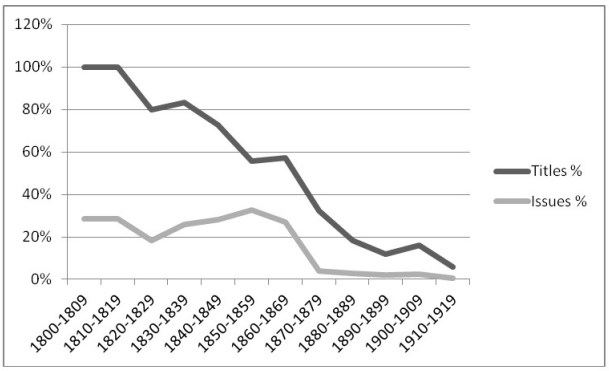

Figure 1: The proportion of titles and issues extracted for the corpus out of the available material in the *Tímarit.is*-archives by decades (1800-1919)

In the early period, text from the great majority of published papers has been extracted and these are around a quarter of available issues. In the late period, on the other hand, text from a small minority of available titles and issues has been extracted. This partly reflects the increase in published newspapers and periodicals in the last 4-5 decades of the period, as the amount of text extracted for the language corpus actually increases considerably over time (cf. Figure 2 below).

The selection of texts from the early period is further limited by the fact that material printed in the only press available in Iceland until about 1840 were in fraktur typesetting, which is practically unreadable by the OCR reading tools currently applied at The National and University Library. The amount of errors in these texts make it unfeasible to try to correct them (semi-)automatically, and most of the earliest texts in the corpus would therefore need to be from Icelandic papers printed and published in Denmark, where latin typesetting became frequent some decades before. To compensate for this, at least to a certain degree, text from a few issues of the earliest papers was entered manually to be included in the corpus.

The texts were selected according to the criteria described above, and extracted from the *Tímarit.is* database in two stages:

1. Texts from ca. 1870-1920
2. Texts from ca. 1817-1870

In addition, two sets of texts were handled separately:

4. Texts from ca. 1800-1840, printed with fraktur typesetting, and entered manually from the images
5. Texts from one periodical, 9 issues, published 1835-1847, which were handled separately in a related project (Daðason et al., 2014)

The selection was done by linguists and lexicographers at The Árni Magnússon Institute for Icelandic Studies, and the technicians at the National and University Library then

[7] A dictionary of 19th and 20th centuries loanwords in Icelandic (Guðrún Kvaran; cf. `http://www.arnastofnun.is/page/tokuord_19_old_20_aldar`).

[8] *Language Change and Linguistic Variation in 19th-Century Icelandic and the Emergence of a National Standard*; cf `http://www.arnastofnun.is/page/LCLV19_project`.

[9] This amounts to a total of approx. 270,000 pages of scanned and OCR read text, but figures for the number of running words could not be obtained.

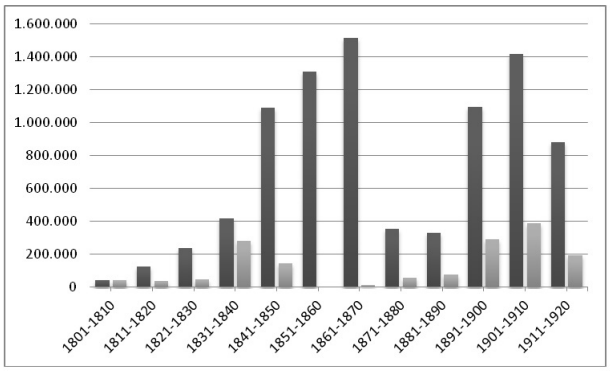

Figure 2: Overview of the amount of texts by decades, total of extracted texts (dark) and already corrected texts (light), measured in number of running words

extracted the selected OCR read texts from the *Tímarit.is*-archives. The selection was based mainly on the type and content of texts, as well as the place and year of publication, according to the criteria described above. The texts were extracted by issue, i.e. each text file contains one issue of a particular newspaper or periodical. The first collection of texts, acquired in 2010, consists of 625 issues of 29 different titles, a total of approx. 4.1 million running words. These texts were the basis for the development of tools for the (semi-)automatic correction, and, as a part of this process, texts from approx. 190 issues of 28 different titles (a total of 1.4 million running words) have been fully corrected, either manually or semi-automatically (cf. section 3.3. below). The second collection includes approx. 560 issues of 14 different papers, again a total of approx. 4.1 million words. These were extracted in 2013, and still await correction. In the third collection, there is manually entered text from the early period, 32 issues (only a part of a few of them) of 7 titles, a total of 256,000 running words. The text entry was done on the basis of the scanned pages displayed at the *Tímarit.is* website. The fourth collection contains the entire text of one particular periodical, published in the 1830s and 1840s (9 issues), a total of approx. 300,000 running words. An overview of the amount of material and its distribution over time, both the extracted text and the texts already corrected or entered manually, is shown in Figure 2.

The figure displays the uneven distribution of the amount of extracted texts over the period, and indicates that there might still be need to add texts from certain decades, and diminish the text material from others, in order to further balance the corpus. It also shows that the amount of texts already prepared is unevenly distributed as well.

3.3. Correcting the OCR read texts

The digitized text from the newspapers and periodicals contains a number of errors that are a result of the optical character recognition (OCR) process. The quality of the OCR reading depends on the quality of the original document, both of the paper and the printing. When lead was used for printing, the printing faces got worn over time, and it also happened that the printing face of individual letters got damaged. In earlier times, when there were few printing presses in Iceland, the same letters were used for a long time, and printed material from the last part of the life of a particular set of letters can be almost unfit for OCR reading. The quality of the paper of old newspapers and periodicals is sometimes very poor and the pages may contain creases that disturb the quality of the OCR. The Icelandic alphabet contains a number of letters with diacritics (*á*, *é*, *í*, *ó*, *ú*, *ý* and *ö*), as well as the letters *ð*, *þ* and *æ*. A review of over one million manually corrected words in a large-scale digitization effort at the Icelandic parliament revealed that over 56% of all word errors involved the misrecognition of one of these characters (Daðason, 2012). Furthermore, the spelling of old texts is not standardized, as there was no agreement on a common standard for Icelandic orthography before the 20th century, and the earliest official standard was only put forth in 1918 (Jónsson, 1959). As previously mentioned (cf. section 3.1.), the correction of texts within the present project aims at a diplomatic version of the original.

When a decision was made to build a corpus based on OCR read texts from the *Tímarit.is* database, it was clear that relying solely on manual correction was infeasible due to the scale of the project. Instead, methods and software for automatically correcting digitization errors would have to be developed in order to compile a corpus of a decent size. An experiment was therefore carried out in 2010, based on the first selection of texts (cf. section 3.2. above). A portion of the selection, 24 issues (in total about 150,000 running words) from various newspapers and periodicals released throughout the time period, were manually corrected with comparison to the scanned images on the *Tímarit.is* website. At the same time, the development of the software, which is based on the principles for spelling correction, was initiated. The software uses frequency information and a lexicon derived from the corrected texts, as well as lists of word forms extracted from lexicographic historical archives and text collections at The Árni Magnússon Institute for Icelandic Studies (Daðason et al., 2014). Based on the manually corrected issues it was estimated that the digitized texts had an average word accuracy of about 91%. The uncorrected versions of the texts were run through the software and the result compared to the manually corrected versions. It was estimated that about 60–65% of digitization errors were corrected with the software. A considerable number of uncorrected issues from the text collection were run through the software, and the remaining errors in these issues were then corrected manually by students. At the end of this phase 51 issues from various newspapers and periodicals, a total of about 290,000 running words, had been corrected.

Previous work (Cushman et al., 1990) has shown that for correction of digitized text to be profitable it needs to have at least 98% character accuracy. If it is assumed that the mean number of characters per word is 5[10] and that character errors are evenly distributed this amounts to 90% word accuracy. Based on these figures it seems to be profitable to correct the digitized text in our case, especially after it has

[10]No figures are available for Icelandic, this is a pure assumption.

been run through the correction software.

Work on the material was continued in 2011, by manual correction of issues that had already been automatically corrected. Halfway through this phase the lexicon used by the software was updated with the vocabulary of the texts corrected so far. Additionally, the spellchecker was supplied with a list of corrections to known word errors, which was derived from the manually corrected texts. After that it was estimated that the software automatically corrected about 77% of all word errors in the scanned texts. The cycle was repeated by running new issues from the text collection through the improved software followed up by manual correction.

In the last phase of the correction project, carried out in 2012, the development of an interactive spellchecker was undertaken. This was meant to replace the process of first running the text through the correction software and then correcting the remaining errors manually. The software (Daðason et al., 2014) was based on the previous work on the automatic correction of digitization errors in our text material, and on other related projects (Daðason, 2012). As well as updating the lexicons the software was supplied with noisy channel functionality for spelling correction (Brill and Moore, 2000). The program, now called *Skrambi*, is a web application, and it accepts uncorrected OCR read text, underlining possible errors and suggesting corrections. The user is given 5 suggestions ordered by probability of being correct, as determined by the noisy channel model. The first suggestion is the correct one in 72% of occurrences, and the correct word is among the top five shown in 84% of occurrences. If none of the suggestions is correct the user can ignore them and make his own correction. An image of the original document is always accessible to the user.

An experiment was performed to test the efficiency of the software. First the time taken to manually correct the errors in 9 issues, each from a different newspaper or periodical from the text collection, was measured. The texts were not run through the correction software first. Then comparable issues (wrt. length and word accuracy) were selected, typically the next issue of the same titles, and these were corrected semi-automatically with the aid of *Skrambi*. The results of the time measurements indicate that correction of the OCR read text is on average three times faster with the help of *Skrambi* than correcting text which has not been run through the software. It was also found that the higher the word accuracy the greater the efficiency of the correction process becomes. This agrees with the results of (Cushman et al., 1990). This is valid for corrections both with and without the help of *Skrambi*. If the OCR read text has high word accuracy the increase in efficiency can be almost fivefold.

3.4. Present state of the project and remaining tasks

To date the text in 188 issues of 28 newspapers and periodicals from the period 1870 to 1920, in total about 1.4 million running words, has been run through the correction process. All issues have been checked manually. The corrected texts are already available for search in the 'Icelandic Text Collection' (*Íslenskt textasafn*) at the website of The Árni Magnússon Institute for Icelandic Studies (under the heading "Blöð_og_tímarit_1860-1930" (Newspapers and periodicals 1860-1930)).[11] The search in this database is limited to strings of letters. The interface allows the user to enter the lemma (base form) of a word (or two adjacent words), and with linking to the 'Database of Modern Icelandic Inflection' (Bjarnadóttir, 2012) the search program seeks out all inflectional forms of the selected word(s), though only with the modern standard spelling. The user can search for all these forms, or deselect forms that he or she chooses to disregard. The texts are useful for various tasks, e.g. in lexicology and lexicography.

For a more focused search in the 19th and early 20th century texts, that could find all possible spelling variants of words and word forms, as well as allow the possibility of searching for grammatical features, a second, standardized layer of tagged texts would be needed, as described in 3.1. above. The next stage in the development of the corpus is therefore to add such information to the texts.

As previously mentioned, the spelling in the newspapers and periodicals, intended for the corpus, differs from the modern spelling norm in various ways, and as the orthography was not standardized in the period, there occur many variants of the same word form. This has consequences both for searching the texts, and for the application of available language technological tools, such as the tools that have been developed for the tagging of Modern Icelandic text (Loftsson, 2008). Rather than adapting the tagging tools to the different spelling variants appearing in Icelandic texts from the 19th and early 20th centuries, we will map the texts to the modern Icelandic spelling standard. The software *Skrambi*, already mentioned in connection with the correction of digitizing errors, will be used for this purpose. *Skrambi* has evolved into a multipurpose software that can be adapted to various tasks, such as spelling correction, the correction of digitized text, and for mapping text between different spelling variants. This is achieved by providing the software with different lexicons. For the mapping between a diplomatic version of 19th century text and a version with modern spelling, a lexicon based on the 'Database for Modern Icelandic Inflection' (Bjarnadóttir, 2012), together with lexicons based on the old texts already corrected, and other language data from the same period, will be used. If time and resources permit, the mapping will be performed semi-automatically with an interactive version of *Skrambi* which works in a similar way to the correction of OCR errors (cf. 3.3 above). After mapping to the modern spelling norm, the texts will be tagged with the system used for tagging the 'Tagged Icelandic Corpus' (*MÍM*) (Helgadóttir et al., 2012). The original, corrected diplomatic text from the newspapers and periodicals, and the tagged text with standardized modern spelling will form a parallel corpus. A search interface will be developed for this corpus using the *Glossa* system (Johannessen et al., 2008) which offers possibilities for search in multilingual corpora. The *Glossa* system has already been adapted to

[11] Cf. `http://corpus.arnastofnun.is/`; there are a few texts from the 1860s and 1920s, which explains the heading.

modern Icelandic for *MÍM*.[12]
Text from early newspapers and periodicals, amounting to about 8.8 million running words, has been extracted from the *Tímarit.is* database in several stages (cf. section 3.2.). Of these, close to 1.6 million running words have been corrected, or in a few cases entered manually. An overview of prepared texts is given in Table 1.

Corpus of early Modern Icelandic			
	Number of titles	Number of issues	Running words
1800-1820	3	7	78,427
1821-1840	5	21	327,357
1841-1860	2	13	146,439
1861-1880	7	29	71,045
1881-1900	17	92	364,124
1901-1920	10	65	579,427
Total	**—**	**227**	**1,566,819**

Table 1: An overview of the amount of texts already prepared for the Corpus of early Modern Icelandic (1800-1920) and their distribution over the period (the texts in the table have either been manually and/or automatically corrected, or manually entered)

A considerable number of texts from the first part of the 19th century has already been prepared, either by running it through the correction procedure (as part of a separate project; (Daðason et al., 2014)) or by manually entering the text (cf. section 3.2.). A few more issues from this early period will be run through the semi-automatic procedure in the next few months. The corrected or manually entered texts from the early part of the period will then be used to adapt the software and increase its effectiveness with respect to the oldest texts. After that the multipurpose program *Skrambi* will be applied to automatically correct the remaining texts and then map them to a standardized modern Icelandic version. After that these texts will be tagged automatically with the available tools. In the absence of any manual corrections, there will surely remain errors in the corrected texts and they will cause further errors in the mapping which again will cause errors in the tagging. It is, however, anticipated that search in these texts will give better results than search in the uncorrected OCR read texts available on the website *Tímarit.is*, and even the corrected but untagged texts in the 'Icelandic Text Collection'. The automatically corrected texts, which are an approximation to a diplomatic version, and the mapped and tagged texts will be a separate part of the intended Corpus of early Modern Icelandic.

4. Conclusions

In the paper, we have described and discussed an attempt to build a corpus of early Modern Icelandic, intended mainly to serve the needs of the linguistic and lexicographical research community, as economically as possible with respect to time and money. We have succeeded to lay the foundations of such a corpus, by using a selection of texts already digitized with OCR methods. This raw material is, however, deficient as it has not been controlled or corrected in any way. The need for a more efficient resource arose in connection to a number of ongoing research projects and the task of constructing a new corpus was undertaken by a group of linguists and specialists in language technology at The Árni Magnússon Institute for Icelandic Studies. Their task was to select the text material, develop methods and tools to correct such material automatically, and at the same time keeping the unstandardized and variant spelling and word forms of the original. This diplomatic text version then needs to be converted to a parallel version with modern standardized text, to enable the application of tools for text analysis, i.e. for automatic tagging of grammatical features and lemmatization. Both versions are then to be presented as a parallel corpus in an easily accessible database with effective and flexible search possibilities. Even if the project is not completely finished, we consider it to be advanced enough to show that the procedures applied have been successful, and we foresee that it will only have taken about four years, with very limited financial resources, to build a useful language corpus with 19th and early 20th century Icelandic.

5. Acknowledgements

Work on the corpus and the development of tools for correcting and mapping text has been a collaboration between linguists, lexicographers, and specialists in language technology at The Árni Magnússon Institute for Icelandic Studies and The University of Iceland, and a number of student assistants have worked on the preparation of texts and the development of methods and tools. Apart from the authors, who have been responsible for data selection and involved in the overall organization and supervision of the project, the following persons should be mentioned: Kristín Bjarnadóttir supervised much of the work on corrections, error analysis and development of methods, esp. in the early stages, as well as organizing an independent project which has enriched the corpus building, both as regards the contents and methods for the automatic mapping of a diplomatic text and a standardized version; Jón Friðrik Daðason started out as a student assistant for developing methods for (semi-)automatic corrections of OCR read texts, but carried on and made use of his experience in the project a strand of his MS-thesis, of which the *Skrambi*-program was a product (Daðason 2012); Kristján Rúnarsson, Hjördís Stefánsdóttir, Guðrún Línberg Guðjónsdóttir and Kristján Friðbjörn Sigurðsson worked on the manual and semi-automatic corrections of the texts at various stages of the project; Örn Hrafnkelsson and Kristinn Sigurðsson at The National and University Library of Iceland have provided the OCR read texts from the *Tímarit.is* archives.
A number of tools and resources used for the construction of the corpus, have been compiled and developed in previous projects. A great number of scholars and students have taken part in these and we want to acknowledge their contribution to the present project, even if it is impossible to list all their names.
The construction of the corpus has been connected to a

[12]Cf. `http://mim.arnastofnun.is/`.

number of ongoing research projects, and part of their funding has gone into the building of this resource: *The strengthening of the text archives at The Árni Magnússon Instistute* (Guðrún Kvaran and Sigrún Helgadóttir) in connection with the compilation of a dictionary of 19th and 20th century loanwords and *Foreign influence in late 19th and 20th century Icelandic* (Ásta Svavarsdóttir), both funded by grants from the University of Iceland Research Fund, as well as the project *Language Change and Linguistic Variation in 19th-Century Icelandic and the Emergence of a National Standard*, supported by the Icelandic Reasearch Fund (grant nr. 120646021/2/3 2012-14; PI: Ásta Svavarsdóttir). Jón Friðrik Daðason got a grant from the the Icelandic Student Innovation Fund 2011 for his program development, and in 2012 the same fund gave support to the project *Fjölnir fyrir hvern mann*, aimed at the correction of a particularily complicated text of one periodical, as well as the development of methods for automatic standardization and mapping between versions. Besides, the corpus construction has benefited from financial support from the Directory of Labour, in the form of summer wages to student assistants working on particular tasks within the project.

6. References

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 – AfLaT, LREC 2012*, pages 13–18, Istanbul, Turkey.

Brill, E. and Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong.

Cushman, W. H., Ojha, P. S., and Daniels, C. M. (1990). Usable OCR: what are the minimum performance requirements? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152.

Daðason, J. F., Bjarnadóttir, K., and Rúnarsson, K. (2014). The Journal *Fjölnir* for Everyone: The Post-Processing of Historical OCR Texts. In *Proceedings of the workshop Language resources and technologies for processing and linking historical documents and archives-Deploying Linked Open Data in Cultural Heritage, LRT4HDA, LREC 2014*, Reykjavík, Iceland.

Daðason, J. F. (2012). Post-Correction of Icelandic OCR Text. MS thesis at University of Iceland, http://hdl.handle.net/1946/12085.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. (2012). The Tagged Icelandic Corpus (MIM). In *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 – AfLaT, LREC 2012*, pages 67–72, Istanbul, Turkey.

Hrafnkelsson, Ö. and Sævarsson, J. (2014). Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books. In *Proceedings of the workshop Language resources and technologies for processing and linking historical documents and archives-Deploying Linked Open Data in Cultural Heritage, LRT4HDA, LREC 2014.*, Reykjavík, Iceland.

Johannessen, J. B., Nygaard, L., Priestley, J., and Nøklestad, A. (2008). Glossa: a Multilingual, Multimodal, Configurable User Interface. In *Proceedings of LREC 2008*, pages 617–621, Marrakesh, Morocco.

Jónsson, J. A. (1959). Ágrip af sögu íslenzkrar stafsetningar. [An overview over the history of Icelandic orthography.]. *Íslenzk tunga/Lingua Islandica*, 1:71–119.

Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1):47–72.

Rögnvaldsson, E. and Helgadóttir, S. (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In Sporleder, C., van den Bosch, A. P. J., and Zervanou, K. A., editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 63–76. Springer, Berlin.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the* 8^{th} *International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey.