The Journal *Fjölnir* for Everyone: The Post-Processing of Historical OCR Texts

Jón Friðrik Daðason, Kristín Bjarnadóttir, Kristján Rúnarsson
The Árni Magnússon Institute for Icelandic Studies
University of Iceland
E-mail: jfd1@hi.is, kristinb@hi.is, krr1@hi.is

Abstract

The journal *Fjölnir* is a much beloved and romanticized 19th century Icelandic journal, published in 1835-1847, which is accessible in digitized form in the digital libraries of the National and University Library of Iceland. In the 19th century, Icelandic spelling was not standardized, and the *Fjölnir* texts were used for spelling experimentation. The spelling is therefore very varied. In the project described in this paper, the aim was making the text of *Fjölnir* accessible on the Web, both in the original spelling, and in modern (standardized) spelling, in a version suitable both for scholars and the general public. The modern version serves two purposes. It makes the text more readable for the general public, and it allows the use of NLP tools made for Modern Icelandic. The post-processing of the OCR texts described in this paper was done with the aid of an interactive spellchecker, based on a noisy channel model. The spellchecker achieved a correction accuracy of up to 71.7% when applied on OCR text, and 84.6% when used to normalize the 19th century text to modern spelling.

Keywords: OCR post-correction, historical texts, spelling normalization

1. Introduction

The topic of this paper is a description of a pilot project making historical Icelandic texts accessible to various groups of users. The texts are from the early 19th century journal *Fjölnir*, which will be made available in the original historical spelling, and in modern Icelandic spelling, with links to digitized copies of the originals.[1] The aim is serving both the general public and researchers, such as historians, linguists, etc., by using NLP tools developed for modern Icelandic, while also providing access to the original text. By adding PoS tagging and lemmatization, both the general public and scholars will be able to access the data in an efficient way. For scholars, the modern version will clearly be a secondary one, i.e., a layer to facilitate analysis, but for the general public the modern version removes the irritation of unfamiliar spelling, sometimes found to be prohibitively annoying. As the morphology of Icelandic is quite rich, and the ambiguity of word forms is extensive, PoS tagging and lemmatization are of great importance, even in the most elementary search (Bjarnadóttir, 2012).

Software for the post-correction of modern OCR text was adapted for this project (Daðason, 2012). The result is a web-based spell checking application based on a noisy channel model, which can be used to achieve a true copy of the original spelling of historical texts, and to produce a parallel text with modern spelling. The software is adapted and used with different lexicons and training data for each task.[2] The project described here is also an attempt at creating the infrastructure for an archive of historical Icelandic texts, where the texts are made accessible to various groups of users, as tailor-made resources are not practicable in a tiny language community.

The organization of the paper is as follows: In chapter 2 the choice of the text for the project is described. Chapter 3 contains comments on Icelandic spelling and language cohesion. Chapter 4 contains the body of the paper, describing the process of post-correction, methodology, the noisy channel model, the error model and the language model. The evaluation of the process is in chapter 5. Chapter 6 shows examples from the Web production. Conclusion and thoughts on the future of the project are to be found in chapter 7.

2. The Journal *Fjölnir*[3]

The Icelandic journal *Fjölnir* was published in 1835–1847, and the instigators were four young Icelandic intellectuals in Copenhagen. The topics of the journal were varied, ranging from articles on politics, history, natural sciences (e.g., ornithology, geology, astronomy, and ichthyology), to articles on language and spelling reforms. Book reviews are an important part of *Fjölnir*, and short stories and poetry are also included, both original Icelandic works and translations. Some of the literary texts are among the most exquisite works of the period, known by most Icelanders. The journal appeared in 9 issues, yearly, with intervals, as shown in Table 1.

Fjölnir was chosen for the project described in this paper partly because of its immediate appeal to the Icelanders, as the journal was both very influential in the Icelandic struggle for independence in the 19th century, and also a

[1] The OCR process itself is not a part of the *Fjölnir* project. The *Fjölnir* texts are a part of the digital library of Icelandic newspaper and journals, *Tímarit.is*, produced and maintained by the National and University Library of Iceland (Hrafnkelsson & Sævarsson, 2014). The OCR post-correction described in this paper is also used in the creation of a corpus of early Modern Icelandic (Ásta Svavarsdóttir et al., 2014).

[2] The spellchecker, named Skrambi, is in fact used for other purposes also, i.e., in context sensitive spellchecking for Modern Icelandic.

[3] *Fjölnir* is one of the names of the Norse god Óðinn (Odin).

cornerstone in the evolution of the romantic period in Icelandic literature. An additional reason for choosing this text is that the spelling of the journal poses interesting problems in itself. At the time of the publication of *Fjölnir*, Icelandic spelling was not standardized, and one of the aims of the four original authors of the journal was establishing very drastic spelling reforms and standardization. These proved to be too drastic to be acceptable to the public, and they were in fact only used partly in the first two issues of the journal. The result is that the spelling of *Fjölnir* is extremely varied, and therefore a real challenge in the OCR post-correction process.

Year	*Pages*	*Words*	*Characters*
1835	180	41,951	243,713
1836	108	31,968	185,994
1837	114	34,272	202,851
1838	92	26,186	155,445
1839	186	59,484	343,139
1843	88	15,974	95,381
1844	140	42,646	248,671
1845	84	20,824	121,975
1847	96	22,867	131,365
Total	1,088	296,172	1,728,534

Table 1: Figures for the journal *Fjölnir*.

3. The Cohesion of Icelandic

The Icelandic language community is very small, with approximately 320 thousand speakers, and limited financial resources. It is therefore imperative to be able to use NLP tools made for modern Icelandic for older Icelandic texts (Svavarsdóttir et al., 2014). Developing NLP tools for each period is too costly to be feasible, even for the periods for which there are sufficient texts for such an undertaking to be remotely possible. As Icelandic spelling was not standardized until modern times, variation has to be taken into account anyway, and the method adopted in this project entails using the modern language to anchor all variants of word forms to lemmas in the Database of Modern Icelandic Inflection (Bjarnadóttir, 2012). This is feasible because the cohesion of Icelandic word forms through the history of the language is sufficiently stable to make the modern forms predictable. In fact, experiments with 15th century texts have shown that approximately 40% of the word forms there were identical to forms in the modern language, after the older texts have been transcribed to a modern character set, without a change of spelling.[4] Because of this, spellchecking methods can be used, and a translation system is not needed.

This does not imply that there have not been linguistic changes through the centuries of Icelandic language history. Part of the motivation of undertaking this project is precisely making the texts available for research on that topic. However, both the rules of word formation and inflection are stable enough and predictable enough for this method to work, as is the vocabulary.

4. The Post-Correction Process

The OCR process has introduced a large number of errors to the text from *Fjölnir*. In this work, we will focus on correcting word errors, to the exclusion of zoning errors where the OCR software has failed to correctly recognize the layout of the text, resulting in text appearing out of order. The zoning errors were corrected manually in this project.

As the same kinds of character recognition errors tend to occur within a given document, a noisy channel model is a good fit, as it can efficiently model the probability of a particular error occurring.

4.1. Previous Work

Tong and Evans (1996) present a method for the correction of OCR errors using a noisy channel model approach combined with a bigram language model. They report an error reduction rate of 60.2% when the method is evaluated on digitized newspaper texts in modern English.

Volk et al. (2011) compare various strategies for reducing OCR errors in a multilingual corpus of digitized 19th century texts. These strategies include enlarging the modern lexicon of the OCR software with words from the targeted time period, applying predefined character substitution rules as well as applying a merging algorithm between the outputs of multiple OCR tools. They combine all correction methods in a single pipeline and find that only the merging algorithm has a significant positive contribution to the overall quality of the text, and that turning it off results in up to a 20% increase in uncorrected OCR errors.

Jurish (2010) generates candidates for the normalization of historical word forms using a variety of methods, including the application of hand-crafted transformational rules and phonetic conflation. The likeliest candidate is chosen using a HMM (Hidden Markov Model) based on a corpus of contemporary German. An F-Score of 99.4% is achieved when this method is applied on a corpus of historical German dating from 1780 to 1880.

Oravecz et al. (2010) normalize historical Hungarian word forms using a noisy channel model combined with a morphological analyzer and a decision tree. The error model is trained on a parallel corpus of manually normalized historical texts. The normalizer achieves a precision of 73.3% when evaluated on Old Hungarian texts.

Bollman et al. (2011) describe a rule-based approach to normalization, where transformational rules are automatically derived from a word-aligned parallel corpus of historical and modern texts. This method increases the ratio of tokens with correct modern spelling from 64.7% to 83.8% when applied on a version of the Bible in historical German. Limiting normalization candidates to word forms which appear in the Bible further improves the ratio to 91.0%.

[4] These experiments were carried out in trials of our normalizer. The texts are a part of the Parsed Historical Icelandic Corpus, IcePaHC: http://www.linguist.is/icelandic_treebank/ (Rögnvaldsson et al., 2012).

Pettersson et al. (2012) normalize a selection of historical Swedish texts using a small number of hand-crafted transformational rules, raising the average number of tokens with modern spelling from 65.2% to 73.0%. Applying contemporary NLP tools on the normalized text was found to yield improved results for a variety of tasks, including verb and complement extraction.

4.2. Methodology

Sufficient language resources for the creation of a lexicon with a reasonable coverage of 19th century Icelandic are available, but the lack of historical corpora precludes the use of statistical language models (beyond unigram models). The problem is compounded by the morphological richness of the language. Also, while Volk et al. (2011) achieved some success by improving the OCR process itself and by utilizing the output of multiple OCR tools, the work on *Fjölnir* is limited to the post-correction of the OCR text.

The OCR process may be likened to transmitting a text string through a noisy channel. The channel may introduce errors to the text by replacing certain characters with others which are similar in appearance. The errors which can occur in a digitized document depend on a number of different factors, such as the OCR software used, the font(s) used, the condition of the paper, and the quality of the scanned image. The probability of a given error, such as the likelihood that the letter *l* could be replaced with an *i*, can vary considerably from one digitized document to another, based on these (and other) factors. The word *ljós* 'light' might therefore consistently be replaced with the nonword *ijós* in one document, yet always be recognized correctly in another. However, even between different documents, it is always unlikely that characters with dissimilar shapes (such as *i* and *s*) be confused.

4.3. Noisy Channel Model

The noisy channel model approach to spelling correction combines an error model and a language model in order to estimate the probability that a misspelled (noisy) string *s* should in fact be the string *w*. The noisy channel model probability is estimated by multiplying the probabilities from the error model and the language model.

4.4. Error Model

The error model estimates the probability that a certain transformation can occur to a string which has been transmitted through the noisy channel, and is trained using pairs of strings prior to and after the transmission.

Kernighan et al. (1990) derive the probability of specific edit operations (the deletion, insertion and substitution of single characters and the transposition of two adjacent characters) from each string pair. The error model probability that *ijós* should be corrected to *ljós* would be calculated as

$$P(ijós|ljós) = \frac{sub(i,l)}{count(l)}$$

where $sub(i,l)$ is the number of times where the letter *i* was replaced with an *l*, and $count(l)$ is the number of times *l* appeared in the training corpus.

This method is improved upon by Brill and Moore (2000), whose model can deal with multiple distinct errors within the same string, while also modeling multiple character edit operations, such as *ph* → *f* in *physical* or *ante* → *anti* in *antechamber*. According to their error model, the probability that the nonword *sern* should be corrected to *sem* 'which' could be calculated as

$$P(sern|sem) = P(s|s) * P(e|e) * P(rn|m)$$

where $P(rn|m)$ is computed as

$$P(rn|m) = \frac{count(m \rightarrow rn)}{count(m)}$$

and $count(m \rightarrow rn)$ is the number of times the letter *m* was replaced with *rn* in the training corpus and $count(m)$ is the number of times *m* occurred in the correct strings. This approach will be followed in this work.

4.5. Language Model

The language model is an n-gram model that returns the probability of a given word. It can be constructed from a lexicon of correctly spelled word forms along with their frequencies (i.e., a unigram model), or derived from some large text corpus. For the purpose of OCR correction, a unigram language model derived from the Database of Modern Icelandic Inflection (DMII; Bjarnadóttir, 2012), which contains approximately 5.8 million Icelandic word forms, along with word frequencies from the 500 million word Web corpus Íslenskur orðasjóður (Hallsteinsdóttir, 2007) is used. Additionally, historical word forms and their word frequencies from the Written Language Archive (WLA), the main historical lexicographic archive at the Árni Magnússon Institute for Icelandic Studies, are used.[5]

4.6. Unsupervised Training of the Error Model

A drawback to the noisy channel model approach is the need for a training corpus for the error model. As mentioned before, OCR error probabilities can vary considerably between different documents, and therefore a single generalized error model will probably not be a good fit for all circumstances.

In this work, we propose a method for the unsupervised training of the error model. Initially, misspelled words are corrected using only the language model probability (i.e., the word frequency of the candidates). Any word form which is not known to the language model is considered to be a misspelling, and is corrected. The error model is then trained using these corrections. With the error model

[5] DMII: http://bin.arnastofnun.is/
Íslenskur orðasjóður: http://wortschatz.uni-leipzig.de/ws_isl/
WLA: http://www.arnastofnun.is/page/ritmalssafn

in place, the misspellings are corrected again using the full noisy channel model probability. The error model is then retrained using the improved corrections. This process is repeated several times. This is essentially an application of the expectation-maximization algorithm (Dempster et al., 1977).

4.7. Candidate Generation

Candidates are generated by the use of a Levenshtein automaton (Schulz & Mihov, 2002), which returns all words $W = \{w_1, w_2, \ldots, w_n\}$ in a lexicon that are within *n* edit operations of a given string *s*. In the first training iteration, the edit operations are limited to single character deletions, insertions and substitutions. In the following iterations, multiple character edit operations (e.g., $rn \rightarrow m$) are also allowed, and are derived from corrections made in the previous iteration. The partition for $P(s|w_i)$ is determined by the edit operations with which the candidate was generated, though it is quite possible for the same candidate to be generated in multiple different ways, in which case all partitions will be ranked.

First, the spellchecker will attempt to generate a list of candidates which are a single edit operation away from the misspelling. If no such words are found, it will attempt to find all words which are within two edit operations of the error. If no candidates can be generated in this manner, the spellchecker will not offer any suggestions to the user, though the word in question will still be underlined as an error.

4.8. Spelling Normalization

The use of predefined transformational rules has been successful when applied to the task of normalizing historical texts (Jurish, 2010; Bollman et al., 2011; Pettersson et al., 2012). However, as the spelling in *Fjölnir* is extremely varied, different rule sets might be needed for each issue. While such rule sets might yield good results when applied to the specific task of normalizing *Fjölnir*, their suitability for normalizing other historical texts, including ones from other time periods, is not as certain. A more general approach is therefore desirable.

As is the case with OCR post-correction, the task of spelling normalization can be viewed as a spellchecking problem. In this sense, historical variants of modern words are considered to be spelling errors that must be corrected to their modern forms. As with OCR texts, the probability of a given transformation can vary wildly between documents.

The noisy channel model is used to normalize *Fjölnir* using the same methods as for the OCR post-correction process, but replacing the historical language model (and lexicon) with a modern one. Here, the language model is derived from the Database of Modern Icelandic Inflection combined with word frequencies from Íslenskur orðasjóður, and the same training method as described above is used to adapt to the characteristics of individual documents.

5. Evaluation

The methods described in this work are evaluated by applying them to the 8th issue of *Fjölnir*, and comparing the results to the already corrected and normalized versions of the text which were manually reviewed for errors. The evaluation extends only to tokens containing at least one alphabetical character. The OCR text was reformatted prior to evaluation in order to eliminate all zoning errors (i.e., instances where the OCR software failed to output the text in the correct order). No other changes were made to the original text.

5.1. OCR Post-Correction

The 8th issue of *Fjölnir* contains a total of 18,714 alphabetical tokens, of which 2,591 were misrecognized during the OCR process (resulting in a word accuracy of 86.2%). The evaluation results can be seen in Table 2.

	Iter. 1	Iter. 2	Iter. 3	Iter. 4
N=1	38.1%	51.6%	52.9%	52.9%
N=5	49.4%	58.1%	57.8%	58.0%

Table 2: Correction suggestion accuracy for OCR errors.

The table above shows the portion of errors where the correct word is the top suggestion (N=1) or among the top five suggestions (N=5), through four iterations of the training algorithm. The correction accuracy of the spellchecker increases substantially after the first iteration, and remains more or less unchanged after the third. The correct word is among the top five suggestions 58% of the time. A review of the remaining errors shows that a considerable portion has been severely misrecognized by the OCR software, containing too many character errors for the correct (or even any) candidate to be generated (e.g., *töflunum* ‘the tables’ → *töfiiiiiuiu* and *varðveiti* ‘preserve’ → *oartwttt*). Further investigation reveals that this is very common for words in Fraktur (Gothic font), which appear with some frequency in this issue.[6] As the spellchecker can only handle two distinct edit operations within a single misspelled word before giving up (even though each operation can correct multiple character errors), it is unable to make a suggestion for the majority of these errors. Repeating the evaluation with words in Fraktur removed from the text yields the following results:

	Iter. 1	Iter. 2	Iter. 3	Iter. 4
N=1	47.9%	65.0%	66.4%	66.6%
N=5	62.0%	72.0%	72.1%	71.7%

Table 3: Correction suggestion accuracy for OCR errors, excluding words in Fraktur.

[6] Words or phrases in Fraktur are quite often interspersed with Roman fonts in *Fjölnir*. This can be seen in Figure 2, where the title of a book in a book review appears in Fraktur, as do direct quotes. The body of the text of *Fjölnir* is in Roman fonts.

As expected, when words in Fraktur are excluded from the evaluation (which raises the word accuracy to 90.0%), the accuracy of the suggestions improves considerably.

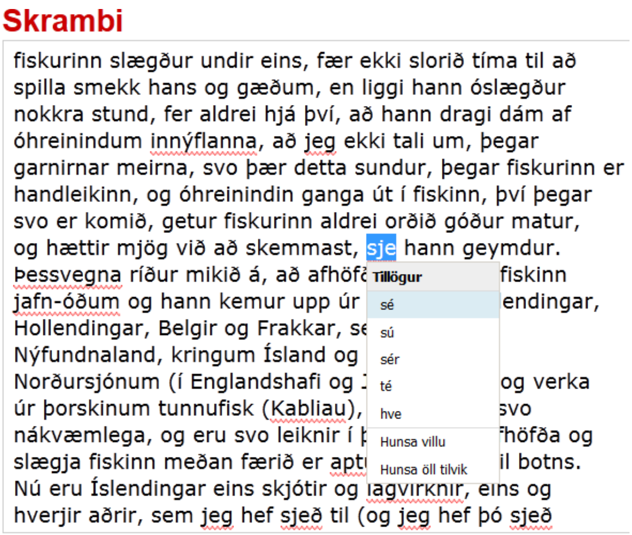
5.2. Spelling Normalization

Applying the noisy channel error model to the corrected version of the text to normalize the spelling to the modern form yields the following results:

	Iter. 1	Iter. 2	Iter. 3	Iter. 4
N=1	35.7%	68.9%	73.4%	73.6%
N=5	48.6%	84.6%	84.6%	84.6%

Table 4: Accuracy of suggestions for the normalization of historical words forms.

The correct modern form is among the top five suggestions in 84.6% of cases. The majority of the remaining errors are real-word errors (most notably where *en* ‘but’ has been written as *enn* ‘still’). These results show that the noisy channel model is well suited to normalizing historical Icelandic text. Figure 1 contains an example of the interactive normalization.



Skrambi

fiskurinn slægður undir eins, fær ekki slorið tíma til að
spilla smekk hans og gæðum, en liggi hann óslægður
nokkra stund, fer aldrei hjá því, að hann dragi dám af
óhreinindum innýflanna, að jeg ekki tali um, þegar
garnirnar meirna, svo þær detta sundur, þegar fiskurinn er
handleikinn, og óhreinindin ganga út í fiskinn, því þegar
svo er komið, getur fiskurinn aldrei orðið góður matur,
og hættir mjög við að skemmast, sje hann geymdur.
Þessvegna ríður mikið á, að afhöfð fiskinn
jafn-óðum og hann kemur upp úr endingar,
Hollendingar, Belgir og Frakkar, se
Nýfundnaland, kringum Ísland og
Norðursjónum (í Englandshafi og I og verka
úr þorskinum tunnufisk (Kabliau), svo
nákvæmlega, og eru svo leiknir í þ fhöfða og
slægja fiskinn meðan færið er aptu il botns.
Nú eru Íslendingar eins skjótir og lagvirknir, eins og
hverjir aðrir, sem jeg hef sjeð til (og jeg hef þó sjeð

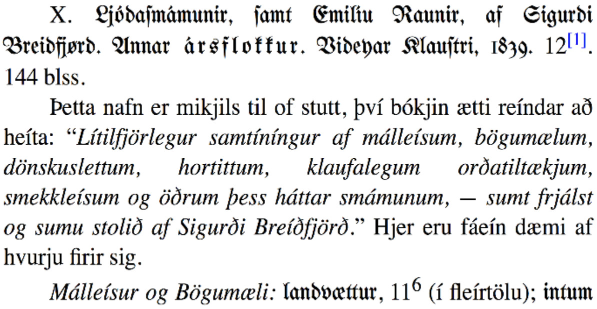
Fig. 1: The interactive spellchecker.

6. Web Production

The goal of the project was to publish *Fjölnir* on the Web, with free access for all users. Following the correction of the OCR text, the final form of the text (in its original spelling) was achieved through manual post-processing and formatting, preserving in the HTML and CSS markup the original layout in a standardized manner: italic, bold, and stretched text, superscript, font face (Fraktur vs. Roman) and size changes, block capitals, headers and subheaders, footnotes, tables, centred poem blocks with left-aligned text, etc. Graphics were drawn up in SVG and MathML was used for fractions and mathematical formulae. Figure 3 contains an example of the original spelling as presented on the website (*Fjölnir* 1843, p. 62); for comparison Figure 2 contains the same text from the original OCR file from Tímarit.is.

XI. £ji5bafmámunir, famt Gnnilíu Sfauntr, af ©ícutrbi 83reíbfj0rb. 2ínnar drðfloffur. S3iber,ar £Iaujtri, 1839. 121. 144 blss. Jietta nafn er niikjils til of stutt, því bókjin ætti reíndar að heífa: "látilffdrlegur smntiningur af málleísum, bögumœlum, dö-nskuslettum, hortittum, klaufalegum orða- tiltækj'um, smekkleisum og öðrum þess húttar smámunum, — sumt frjálst og sumu stolið af Siguroi Breíðfj'úrð." Hjcr eru fáei'n dæmi af hvurju firir sig. Málleísur og Bögumœli. lanbttcettur, ll6 (í fleírtölu); intum

Fig. 2: Example of OCR text, from Tímarit.is.



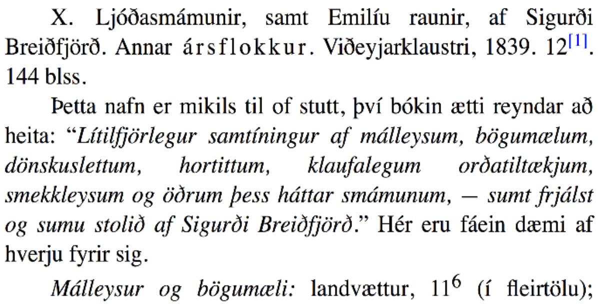
X. Ljóðaſmámunir, ſamt Emilíu Raunir, af Sigurði Breiðfjørð. Annar ársflokkur. Viðeyar Klauſtri, 1839. 12[1]. 144 blss.

Þetta nafn er mikjils til of stutt, því bókjin ætti reíndar að heíta: “*Lítilfjörlegur samtíníngur af málleísum, bögumælum, dönskuslettum, hortittum, klaufalegum orðatiltækjum, smekkleísum og öðrum þess háttar smámunum, — sumt frjálst og sumu stolið af Sigurði Breíðfjörð*.” Hjer eru fáeín dæmi af hvurju firir sig.

Málleísur og Bögumæli: landvættur, 11^{6} (í fleírtölu); intum

Fig. 3: Example from the website, original spelling.[7]

The next step is to create the modern spelling layer, which will inherit the formatting already present in the original spelling layer. The Web interface will allow the user to easily switch between the layers. Figure 4 shows the same text as in Figures 2 and 3, in the modern spelling.



X. Ljóðasmámunir, samt Emilíu raunir, af Sigurði Breiðfjörð. Annar ársflokkur. Viðeyjarklaustri, 1839. 12[1]. 144 blss.

Þetta nafn er mikils til of stutt, því bókin ætti reyndar að heita: “*Lítilfjörlegur samtíningur af málleysum, bögumælum, dönskuslettum, hortittum, klaufalegum orðatiltækjum, smekkleysum og öðrum þess háttar smámunum, — sumt frjálst og sumu stolið af Sigurði Breiðfjörð*.” Hér eru fáein dæmi af hverju fyrir sig.

Málleysur og bögumæli: landvættur, 11^{6} (í fleirtölu);

Fig. 4: Example from the website, modern spelling.

[7] Translation: Poetic Trivia, with the Lament of Emilia, by Sigurður Breiðfjörð. Second annual part. Viðey Monastery, 1839. 12[1]. 144 pps. This title is much too short, because the book should be called “A trivial hotchpotch of blunders, solecisms, Danishisms, waffle, clumsy phrasing, bad taste, and other trivialities, some freely available and some stolen by Sigurður Breiðfjörð.” Here are some examples of each of those. Blunders and solecisms. *landvættur* 11-6 (in the plural); *intum*

Further along, a unified file format incorporating any number of named text layers is envisioned, from which HTML-files in each version may be generated. An example of the information contained in such a unified file may be seen in Table 5.

OCR	*Post-corr.*	*Modern*	*Lemma*	*Tag*
Hjcr	Hjer	Hér	hér	aa
eru	eru	eru	vera	sfg3fn
fáei´n	fáeín	fáein	fáeinir	fohfn
dæmi	dæmi	dæmi	dæmi	sþghfn
af	af	af	af	aþ
hvurju	hvurju	hverju	hver	fsheþ
firir	firir	fyrir	fyrir	ao
sig	sig	sig	sig	fphfo

Table 5: Example of the 3 layers of *Fjölnir* 1838, with lemmas and tags.[8]

The *Fjölnir* website will be accessible from the website of the Árni Magnússon Institute for Icelandic Studies, http://arnastofnun.is/.

7. Conclusion

The two versions of the texts of the *Fjölnir* project are due to be made accessible online in the spring of 2014. A unified file format incorporating the text layers, with annotation, are a part of larger prospective project at the Árni Magnússon Institute for Icelandic Studies.

The noisy channel model proved to be successful, even for a very error-prone OCR text, but using a more complex language model, such as a bi- or tri-gram model would likely improve the correction accuracy for both OCR post-correction and normalization. For better results, context-sensitive error correction is needed for real-word errors. Additional updates to the spellchecker are planned, such as dynamically updating the error model probabilities as the user makes corrections.

While the tool described in this work is interactive, it could easily be converted into a fully automated spellchecker for the correction (as well normalization) of large-scale digitization efforts.

8. Acknowledgements

The project described in this paper has been run in tandem with the creation of a corpus of early Modern Icelandic (Svavarsdóttir et al., 2014), with close cooperation of participants in both projects. We would like to thank Ásta Svavarsdóttir, Sigrún Helgadóttir, and all the participants in the corpus building. Thanks are also due to Örn Hrafnkelsson and Kristinn Sigurðsson at the National and University Library of Iceland for access to the OCR texts of *Fjölnir*. The Icelandic Student Innovation Fund financed the work of Jón Friðrik Daðason and Kristján Rúnarsson on the *Fjölnir* project in the summer of 2012, and the creation of the software was partly financed by a grant from the same fund to Jón Friðrik Daðason in 2011.

9. 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*, pp. 13-18, Istanbul, Turkey.

Bollmann, M., Petran, F., & Dipper, S. (2011). Rule-based normalization of historical texts. In *Proceedings of the International Workshop on Language Technologies for Digital Humanities and Cultural Heritage* (pp. 34-42).

Brill, E., & 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.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1), 1-38.

Hallsteinsdóttir, E., Eckart, T., Biemann, C., Quasthoff, U., & Richter, M. (2007). Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In *Proceedings of NODALIDA-07*, Tartu, Estonia

Hrafnkelsson, Ö., & 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.

Jurish, B. (2010). More than Words: Using Token Context to Improve Canonicalization of Historical German. *JLCL*, 25(1), 23-39.

Kernighan, M. D., Church, K. W., & Gale, W. A. (1990). A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on Computational linguistics-Volume 2* (pp. 205-210). Association for Computational Linguistics.

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

Oravecz, C., Sass, B., & Simon, E. (2010). Semi-automatic normalization of Old Hungarian codices. In *Proceedings of the ECAI Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (pp. 55-59).

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

[8] The text was tagged using IceNLP (Loftsson, 2008), http://nlp.cs.ru.is/icenlp/?lang=en.

Pettersson, E., Megyesi, B., & Nivre, J. (2012). Rule-Based Normalisation of Historical Text–a Diachronic Study. In *Empirical Methods in Natural Language Processing: Proceedings of the Conference on Natural Language Processing* (pp. 333-341).

Schulz, K. U., & Mihov, S. (2002). Fast string correction with Levenshtein automata. *International Journal on Document Analysis and Recognition*, *5*(1), 67-85.

Svavarsdóttir, Á., Helgadóttir, S., & Kvaran, G. (2014). Language resources for early Modern Icelandic. 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.

Tong, X., & Evans, D. A. (1996). A statistical approach to automatic OCR error correction in context. In *Proceedings of the fourth workshop on very large corpora* (pp. 88-100).

Volk, M., Furrer, L., & Sennrich, R. (2011). Strategies for reducing and correcting OCR errors. In *Language Technology for Cultural Heritage* (pp. 3-22). Springer Berlin Heidelberg.