Eiríkur Rögnvaldsson

Old languages, new technologies: The case of Icelandic

1. Introduction

The focus of this workshop is on language resources and tools for processing and linking historical documents and archives. In the past few years, interest in developing language technology tools and resources for historical languages or older stages of modern languages has grown considerably, and a number of experiments in adapting existing language technology tools and resources to older variants of the languages in question have been made. But why should we want such resources and tools? What is the purpose?

The reasons for this increased interest can vary. One is that more and more historical texts are becoming available in digital format and thus amenable to language technology work. As a result, researchers from many disciplines are starting to realize that they could benefit from being able to search these texts and analyze them with all sorts of language technology tools.

As for myself, I need resources and tools for studying diachronic syntax, because that is – or used to be – my main field of research. I started out as a syntactician and have never regarded myself as a specialist in language technology in any way. In the 1980s I wrote several papers on Modern Icelandic syntax but was not particularly interested in historical syntax, historical documents, cultural heritage or anything like that. However, I happened to be interested in computers and bought one already in 1983 – I still keep it in my office.

Since the 1970s, generative syntacticians have shown a continually growing interest in historical syntax and syntactic change. There are two reasons for this. One is interest in language change in general – the desire to know how languages change. Another reason is the comparative or even micro-comparative aspect of generative syntax, where closely related languages are compared in search of parametric variation. The first Diachronic Generative Syntax Conference was held in 1990. In the beginning, this conference was held biannually, but since 2008 it has been held every year, demonstrating the growing interest for the subject.

The nature of the talks at these conferences has also changed. In the beginning, most of the talks were rather theory-oriented, using data from handbooks and other secondary sources to support the argumentation. Many of the speakers had not looked at the primary sources and their interpretation of their data was sometimes doubtful or plainly wrong, due to their lack of philological skill. I remember attending one such conference in the 1990s where I was almost the only speaker who was using examples taken directly from the primary sources – not because I particularly wanted to look at the sources, but because I had no other option – there were no useful handbooks and reference works on historical and diachronic Icelandic syntax.

In recent years, the focus of generative historical syntacticians has shifted towards the use of corpora and statistical methods. People have realized that data from handbooks can be misleading. The handbooks necessarily show only a small proportion of the data available in the primary sources – and that proportion is not randomly chosen, but may rather be claimed to be eccentrically chosen. The examples are selected by the authors of these handbooks and we know that a number of external features may have affected their choice of examples – their background, their motivation, their theoretical assumptions, etc. Furthermore, the handbooks are full of subjective claims like “common”, “rare”, “more frequent than”, etc. We also know that the authors may have – and almost certainly will have – missed some important and even crucial examples.

For these reasons, many historical syntacticians are now busy building corpora and the number of available parsed historical corpora or treebanks is growing fast. However, building parsed corpora for dead languages or older stages of living languages is not a trivial task. First, the texts need to be digitized, and then, they need to be parsed. But usually, no parsers exist for these language stages, and existing parsers for the modern languages cannot be used since the differences between the different stages are that great.

Furthermore, the texts are not normalized – they are usually full of inconsistencies in spelling, morphology, word-forms and syntax. Thus, it is very difficult to build tools for tagging and parsing these texts, or to adapt existing tools to them. Manual parsing can then be the only option. However, even if the texts are parsed manually, we may still want to use some tools for searching them, and that can be complicated if they are full of inconsistencies. Therefore, some sort of standardization is usually necessary at some point.

Another reason for an increased interest in developing language technology tools and resources for historical languages is that since historical texts often exhibit considerable variation in spelling and morphology, they pose great challenges to existing language technology tools and methods developed for modern standardized texts. Thus, many language technology researchers see historical texts as a good test bed for developing and enhancing their methods and tools.

In many ways, Icelandic is well suited for being such a test bed. Icelandic has changed less during the last thousand years than most or all other languages with a recorded history. Icelandic has a relatively large corpus of texts from different stages in its recorded history, starting with a number of narrative texts from the 13th and 14th centuries. There is a strong community of philologists and linguists working on medieval Icelandic, and a number of projects on digitizing and analyzing Old Icelandic texts are ongoing, in preparation, or have just been finished.

Since serious work on Icelandic language technology started around the turn of the century, several important resources and tools for Modern Icelandic have been built, such as the Database of Modern Icelandic Inflection (Bjarnadóttir 2012), the Tagged Icelandic Corpus (Helgadóttir et al. 2012), and the IceNLP package including a POS tagger (Loftsson 2008), a shallow parser (Loftsson and Rögnvaldsson 2007) and a lemmatizer (Ingason et al. 2008), to name the most important ones.

As a result of the META-NORD project (Rögnvaldsson et al. 2012a), most of the existing tools and resources are now open and free for everyone to use under standard licenses (GNU and Creative Commons), and can be downloaded either from a META-SHARE repository or through the website `málföng.is` (Helgadóttir and Rögnvaldsson 2013)

In this talk, the focus will be on resources and tools that have been developed for older stages of Icelandic. I will start by briefly describing Old Icelandic, the main differences between Old and Modern Icelandic, and our main textual sources for Old Icelandic. Then I will talk about the lemmatized Concordance to the Icelandic Family Sagas which was published on CD-ROM in 1996 (Kristjánsdóttir et al. 1996).

After that, I talk about experiments in tagging Old Icelandic with tools developed for or trained on Modern Icelandic texts. Then I talk about experiments in using the rule-based IceNLP package in preparing electronic editions of Old Icelandic texts. After that, I talk about the use of IceNLP in preparing the Icelandic Parsed Historical Corpus, IcePaHC. Finally, I try to link different things together.

Of course, I will not be able to do justice to all experiments or projects which have used tools or resources developed for Modern Icelandic in analyzing older stages of the language, but

two such projects must be mentioned. One is *Greinir skáldskapar* – a tagged, parsed and lemmatized corpus of historical Icelandic poetry, based on dependency syntax (Eythórsson, Karlsson and Sigurðardóttir 2014). The other is the OCR correction of the text of the journal *Fjölnir* (Daðason, Bjarnadóttir and Rúnarsson 2014). Both of these projects will be described in special talks at this workshop later today, so I don't need to say more about them.

2. Old Icelandic

Icelandic is a North Germanic language, which the Norwegian settlers brought with them when they came to Iceland in the late 9th and the 10th centuries (Karlsson 2004). In the beginning, the language spoken in Iceland was of course not different from Norwegian, and in the middle of the 12th century, the difference was still dialectal. However, the dialects gradually drifted apart from each other and in the middle of the 14th century, the difference seems to have become considerable. From then on, it is natural to talk about Icelandic and Norwegian as two separate languages instead of two dialects of the same language.

Icelandic is a language with a rich literary heritage ranging from the 12th century to the present. According to the main source on the earliest history of Icelandic, The Book of Icelanders by Ari the Learned, Icelanders started to write in 1117-1118. The oldest preserved manuscripts are from the latter half of the 12th century. These manuscripts were written on vellum – skin of calf. Most of them contain religious material, translated from Latin.

Around the middle of the 13th century, Icelanders started to write narratives. The narrative texts from the 13th century can be divided into three main categories or genres. One is the so-called Contemporary Sagas – accounts of events in Iceland in the 13th century, the most important of which are *Sturlunga Saga* and *Sagas of Bishops*. Another category is *Sagas of the Kings of Norway*, the most famous of which is *Heimskringla*, a collection of sagas by Snorri Sturluson. The third category is the Icelandic Family Sagas, Íslendingasögur – narratives of people and events in the 10th and early 11th centuries.

In the 14th century, the writing of narratives continued, and two new genres were added which flourished in the 14th and 15th centuries. These are the Legendary Sagas which are fictitious texts which are set before the settlement of Iceland, and Sagas of Knights which started out as translations of the French chansons de geste but later developed into indigenous writings in a similar style.

The Family Sagas are the most famous of these texts. The Sagas are around 40, but many of them exist in two or more variants, sometimes widely different. Around 10 years ago, a new English translation of all the Sagas was published, and a new translation of the Sagas into Danish, Norwegian, and Swedish has just been published. However, the Sagas have never been translated into Modern Icelandic – we'll come to that later.

Like in most other languages, historical texts in Icelandic show great variation in spelling, even though it may be mentioned that in the 12th century, shortly after Icelanders started writing in Latin letters, an unknown person usually referred to as the First Grammarian made an attempt to standardize Icelandic spelling in a famous essay called the First Grammatical Treatise (Benediktsson 1972). But this attempt was not successful.

It must be emphasized that no Icelandic narratives from the 13th and 14th centuries are preserved in the original; the texts are mostly preserved in vellum manuscripts from the 13th through the 15th centuries, but some of them only exist in paper manuscripts from the 16th and 17th centuries. This makes it extremely difficult to assess the validity of these texts as linguistic evidence, since it is often impossible to know whether a certain feature of the preserved

text stems from the original or from the scribe of the preserved copy, or perhaps from the scribe of an intermediate link between the original and the preserved manuscript.

It is well known that scribes often did not retain the spelling of the original when they made copies; instead, they used the spelling that they were used to. In many cases, two or more manuscripts of the same text are preserved, and usually they differ to a greater or lesser extent. Furthermore, it is known that not all editions of Old Icelandic texts are sufficiently accurate (cf., for instance, Degnbol 1985).

3. Old vs. Modern Icelandic

It is a commonly accepted fact that Icelandic morphosyntax has changed much less during the last thousand years than most other European languages. This has often been attributed to the strong literary tradition and the isolation of the country. However, it must be emphasized that some features of the language have in fact changed considerably since Old Icelandic. Thus, the phonological system has undergone dramatic changes, especially the vowel system.

Fortunately, we have very detailed information on the sound system of 12th century Icelandic in the above-mentioned First Grammatical Treatise, probably written around 1140 but preserved in a 13th century manuscript. As already mentioned, the aim of the author was to develop a spelling standard for Icelandic. He points out that the Latin alphabet is not sufficient – it lacks characters for some Icelandic sound (Benediktsson 1972). Then he proceeds to argue which sounds need special characters to denote them.

In his argumentation he uses minimal pairs – a concept which is usually attributed to the Prague School structuralists in the 1930s. He demonstrates that there is a distinctive opposition between long and short vowels – and moreover, in the long vowel system, there is a distinctive opposition between nasal vs. non-nasal vowels. The nasal vowels seem to have been lost in the 13th century, so the nasal vs. non-nasal opposition disappeared. The long vs. short vowel opposition survived until the 16th or 17th century. In Modern Icelandic, vowels can be both long and short, but the distribution is now complementary.

Several other changes have occurred in the phonological system. For instance, the characters *p*, *t*, *k*, no longer denote voiced sounds – all Icelandic stops are unvoiced. Before aspirated stops, sonorants have been devoiced in the majority dialect. Geminate *pp*, *tt*, *kk*, have turned into pre-aspirated stops.

Most of the phonetic changes have one important thing in common: They do not show in the spelling since they do not have a neutralizing effect. An accent mark over a vowel symbol used to denote length, but now it denotes another sound quality than the unaccented symbol. In Old Icelandic, the opposition between /b, d, g/ vs. /p, t, k/ presumably was between voiced vs. voiceless sounds, whereas in Modern Icelandic, it is between unaspirated vs. aspirated sounds. In Old Icelandic, the opposition between geminate /bb, dd, gg/ vs. /pp, tt, kk/ presumably was also between voiced vs. voiceless sounds, whereas in Modern Icelandic, it is between unaspirated vs. pre-aspirated sounds. And so on.

Of course, a number of phonological changes have left their marks on the orthography but these are fairly regular. Among those we can mention the spirantization of stops in final position in unstressed syllables; *ok* > *og*, *þat* > *það*, etc. Admittedly, a few vowel phonemes have merged. Some of the mergers are shown in the spelling – ‘æ’ is now used for both ‘ae ligature’ (æ) and ‘oe ligature’ (œ), and ‘ö’ is used for both ‘o with a slash’ (ø) and ‘o with a hook’ (o̜). However, the merger of /i/ and /y/ is not shown – *y* is still retained in Modern Icelandic spelling even if it has denoted the same sound as *i* for four hundred years or so.

What is noteworthy here is that most of the sound changes involve changes of features – changes in sound quality, in nasality, in length, in voicing, in aspiration, etc. They do not involve massive assimilations, syncope, apocope, or other types of changes which in turn might have led to merger or disappearance of inflectional endings. And in fact, the inflectional system is for the most part intact. Modern Icelandic retains almost all the inflectional categories of Old Icelandic. The only exception is that Old Icelandic still had the dual in personal and possessive pronouns, but it has disappeared from Modern Icelandic.

Furthermore, most of the inflectional classes that we find in Old Icelandic still exist – only one small inflectional class of nouns has merged with a larger class. A number of words have shifted inflectional class but the proportion of those words is very low. In the verbal inflection, the vowel of a few subjunctive endings has changed, in most cases leading to merger of the subjunctive and the indicative, even though one of these changes has the opposite effect.

The sound changes that have occurred in Icelandic are not necessarily fewer or less drastic than those in the Mainland Scandinavian languages, but they are mainly of a different kind, as explained above. It is well known that the inflectional system of the Mainland Scandinavian languages has been greatly simplified. There is of course a complex interplay between inflection and syntax, and it is often claimed that the loss of inflection in the Mainland Scandinavian languages has led to several changes in the syntax, especially as regards word order. The exact nature of this interplay is debated, but there is hardly any doubt that there is some connection here.

Since the sound changes do not have a drastic effect on the spelling, and the inflectional system remains more or less the same, it is possible to claim that Old and Modern Icelandic are essentially the same language. Since the morphological categories are almost the same, we can use the tagset developed for Modern Icelandic to tag Old Icelandic texts with only a minor modification.

The vocabulary has also been rather stable. Of course, a great number of new words (loanwords, derived words and compounds) have entered the language, but the majority of the Old Norse vocabulary is still in use in Modern Icelandic, even though many words are confined to more formal styles and may have an archaic flavor. Since the changes are mainly due to new words being added rather than to old words becoming obsolete, these changes do not pose problems for the adaptation of language technology tools to older stages of the language.

The syntax is also basically the same, although a number of changes have occurred (cf. Faarlund 2004; Rögnvaldsson 2005). These changes involve for instance word order, especially within the verb phrase, the use of phonologically "empty" NPs in subject (and object) position, the introduction of the expletive *það* 'it, there', the development of new modal constructions such as *vera að* 'be in the process of' and *vera búinn að* 'have done/ finished', etc.

Thus, present-day Icelanders can read many texts from the 13th century without special training, although that doesn't necessarily mean that they can read the texts directly from the manuscripts. There was no accepted spelling standard until the 20th century, and the same sounds, sound combinations and words can be written in many ways. However, since the morphology is almost the same, it is usually relatively straightforward to convert older spelling to the modern standard and get legible text.

4. Textual sources

What kind of texts do we have access to? Editions of medieval Icelandic texts are of four different types: facsimile (very close transcription), diplomatic (less close and with some inter-

pretation – abbreviations written out in full), normalized (regularized orthography) and modernized (Modern Icelandic orthography). Each type of editions serves a special group, from graphologists to the general public.

Up to the beginning of the 19th century, everyone used his or her own spelling, usually reflecting a mixture of their own pronunciation and the spelling of the manuscripts they had been exposed to, accompanied by considerable intra-scribal variation. With the advent of periodicals around 1800, and especially after the advent of weekly newspapers around 1850, the spelling gradually became more and more uniform and around 1900, a commonly agreed standard had emerged. Texts from the 19th century onwards usually only have minor deviations from the modern spelling.

In the 19th century, scholars who edited the Sagas developed a normalized orthography which was used in most editions of Old Icelandic texts up to the late 20th century – and is still in use. This orthography is based on the phonological system of Icelandic around 1200. Since many of the sound changes that have occurred during the last 800 years are not reflected in Modern Icelandic spelling anyway, as mentioned above, the differences between the old normalized orthography and Modern Icelandic orthography are not that great.

For most of the 20th century, editions of medieval texts intended for the public were usually in the normalized Old Norse spelling. In fact, this spelling was considered an integral part of the Icelandic cultural heritage and even sacred in some sense. In 1941, the famous writer and Nobel Prize winner Halldór Laxness published one of the Icelandic Family Sagas using Modern Icelandic spelling instead of the old normalized spelling. This caused great uproar and the Parliament even passed laws which prohibited publishing Old Icelandic texts using modern spelling. However, the Supreme Court ruled these laws unconstitutional.

In 1985, a group of young philologists and literary scholars embarked on a large project: A new edition of the Icelandic Family Sagas in Modern Icelandic spelling (Halldórsson et al. 1985-86). Personal computers which were quite new at that time were used in the preparation of this edition but there were no plans of an electronic edition – CD-ROM and the World Wide Web lay somewhere in the future. Around 1990, other main Old Icelandic narrative text collections were published in Modern Icelandic spelling. This includes both *Sturlunga Saga* (Kristjánsdóttir et al. 1988) and *Heimskringla* (Kristjánsdóttir et al. 1991).

In these editions, the text was normalized to Modern Icelandic spelling. This involves, for instance, reducing the number of vowel symbols reflecting mergers in the vowel system as mentioned above, inserting *u* between a consonant and a word-final *r* (*maðr* 'man' > *maður*), shortening word-final *ss* and *rr* (*íss* 'ice' > *ís*, *herr* 'army' > *her*), changing word-final *t* and *k* in unstressed syllables to *ð* and *g*, respectively (*þat* 'it' > *það*, *ok* 'and' > *og*), etc. Furthermore, a few inflectional endings are changed to Modern Icelandic form. It must be emphasized that these normalizations do not in any way simplify the inflectional system or lead to the loss of morphological distinctions in the texts.

A number of texts have also been scanned. This includes the Legendary Sagas (Fornaldarsögur Norðurlanda) which were scanned after editions in normalized Old Icelandic orthography. The OCR scanned text has subsequently been corrected and the texts are available on the web and can be downloaded.

Since 2001, there exists a network of Nordic archives, libraries and research institutions working with medieval texts and manuscripts. This archive is called Medieval Nordic Text Archive, *Menota* for short (`menota.org`), and its aim is to preserve and publish medieval texts in digital form and to adapt and develop encoding standards necessary for this work. The participants in this network come from Norway, Iceland, Denmark and Sweden.

Most of the texts in the archive are in Old Icelandic, but there are also a number of Old Norwegian and Old Swedish texts. In most cases, only one or two versions of each text exist in the archive, but a few texts exist in three different versions. A few of the texts have been lemmatized.

5. Concordance to the Icelandic Family Sagas

In 1987, I became involved with the group of scholars that worked on the above-mentioned editions of medieval Icelandic narrative texts in Modern Icelandic spelling. The reason was that my wife joined this group which already included a number of my friends. Given my interests in computers, I soon started to speculate what could be done with these texts, in addition to publishing them on paper (Rögnvaldsson 1990).

The Institute of Lexicography at the University of Iceland had at that time acquired a copy of a program called BYU Concordance, which later became known under the name of WordCruncher. We started experimenting with this program and found out that by making simple KWIC concordances, a new world opened to us. All kinds of patterns, formulas etc. which had gone by unnoticed suddenly jumped to the eye (Rögnvaldsson 1997).

However, we soon realized that a simple KWIC concordance could not fulfill our needs. Since Icelandic is a heavily inflected language, it is common that different inflectional forms of a word end up at different places in an alphabetically sorted concordance. Furthermore, due to homophony in inflectional endings, homophony of inflectional forms belonging to different words is very common. We thus decided that we had to make a lemmatized concordance.

We started out by producing a KWIC concordance of the whole text, around one million tokens, but the rest of the work had to be done manually. We neither had the necessary programs nor programming skills to perform automatic lemmatization. The only things we could do to speed up the work a bit was to create a number of WordPerfect macros to copy and insert lines or bulks of text. Thus, 1989 was the year I spent lemmatizing one million lines of text. I'm not particularly keen on spending another year the same way.

In November 1989, the lemmatized concordance was almost finished, even though massive proof-reading was still to be done. However, we could start producing results. For instance, we had for the first time information on the size of the vocabulary of the Sagas, the size of different parts of speech, the difference in vocabulary and usage between different Sagas, etc. We were also able to discover subtle semantic nuances which had gone by unnoticed. One problem for Icelanders today who are reading the Sagas is the great similarity of Old and Modern Icelandic.

Our original plan was to publish the concordance as a book – at that time, we didn't realize that there were any other possibilities, and perhaps there weren't in 1990. However, it took some time to clean up the concordance, finish the proof-reading and so on, so when it was finally ready for publication, a CD-ROM edition had become a realistic option and the concordance was published on a CD in 1996 (Kristjánsdóttir et al. 1996). Some ten years later it was put on the Internet and is now accessible online – unfortunately you have to pay for access but that is beyond our control.

In the 1990s the concordance was used by myself and others in various studies of the vocabulary and word usage in the Sagas – studies which would have been impossible without the concordance (Rögnvaldsson 1995; 1996a; 1996b; 2000). However, my work with the texts made me want to utilize my background in syntax in analyzing these texts syntactically, so I shifted from synchronic to diachronic syntax. Of course, I only had the raw texts and the con-

cordance to work with – the texts were neither tagged nor parsed, so I just had to use simple searches for individual words and combinations of words.

6. Tagging Old Icelandic

In the following, I will describe briefly the experiments that have been done in tagging and parsing older stages of Icelandic. In 2001, a group of four Icelandic scholars from four different disciplines – linguistics, lexicography, statistics, and engineering – started to work on language technology, after having attended courses in the then just established Swedish GSLT – Graduate School of Language Technology.

We soon discovered that fortunately, the source files from the *Icelandic Frequency Dictionary* which was published in 1991 (Pind et al. 1991) had been preserved on the server at the Institute of Lexicography. This was by no means self-evident – around 1990, disk space was still rather expensive and many files which we would now want to have were deleted in order to save space – or else transferred to some media which are now unreadable.

The Frequency Dictionary was based on fragments of 100 texts from five different genres, each fragment around 2000 tokens. The files consist of these 500,000 tokens and a tag string following each token. It turned out that these files were ideal for use as training and testing material for a PoS tagger. It turned out that these files were ideal for use as training and testing material for a PoS tagger. From 2002-2004, we tested and trained several different data-driven taggers on the Frequency Dictionary files. The TnT tagger gave the best results, reaching 90.4% accuracy (Helgadóttir 2005; 2007).

The tag strings are analytic, each character denoting a specific value of a morphological category such as case, number, gender, tense, mood, person etc. This tagset was devised for the Frequency Dictionary and has become a de facto standard tagset for tagging Icelandic. Due to the inflectional nature of the language it is rather large, comprising around 700 different tags.

Having trained the TnT tagger on Modern Icelandic texts, we wanted to find out whether the tagger could be of help in tagging Old Icelandic narrative texts, with the purpose of facilitating the use of these texts in research on syntactic variation and change (Rögnvaldsson and Helgadóttir 2011). At a first glance, it may seem unlikely that a tagger trained on 20th century language could be applied to 600-700 years old texts. Nevertheless, we found it worthwhile to try to adapt the tagging model that we had trained for Modern Icelandic to our Old Icelandic electronic corpus.

Our motive was not to get a 100% correct tagging of the Old Icelandic texts, but rather to facilitate the use of the texts in syntactic research. To create a manually annotated training corpus for Old Norse from scratch would have been a very time-consuming task. Thus, the possibility of using bootstrapping methods was a key factor in realizing this project. The whole process took only a small fragment of the time it would have taken to create a manually corrected corpus to train the parsers.

We started by running TnT on the whole Old Icelandic corpus using the tagging model developed for Modern Icelandic (cf. Helgadóttir, 2005; 2007). We then measured the accuracy by taking four samples of 1,000 words each from different texts in the corpus – one from the *Family Sagas*, one from *Heimskringla*, and two from *Sturlunga Saga* – and checking them manually. Counting the correct tags in these samples gave 88.0% correct tags, compared to 90.4% for Modern Icelandic.

Even though these results were worse than those we got for Modern Icelandic, we considered them surprisingly good. The syntax of Old Icelandic differs from Modern Icelandic syntax in

many ways as mentioned above, and one would especially expect the differences in word order to greatly affect the performance of a trigram based tagger like TnT. However, sentences in the Old Icelandic corpus are often rather short, which may make them easier to analyze than the longer sentences of Modern Icelandic.

We then selected seven whole texts (sagas) and two fragments from the *Sturlunga* collection for manual correction – around 95,000 words in all. This amounts to one third of the *Sturlunga* collection. The manual correction was a time-consuming task, but the time and effort spent on checking and correcting the output of TnT was only a small fragment of the time and effort it would have taken to tag the raw text.

We trained TnT on the corrected text, tagged the whole corpus again with the resulting model, and measured the accuracy on the same four samples of 1,000 words each as in the first experiment. Now the results were much better – 91.7% correct tags, which is better than the 90.4% accuracy that we got for Modern Icelandic. It may seem surprising how much the accuracy improved when we used this model, especially when we consider that the training corpus was much smaller than the training corpus for Modern Icelandic (95,000 words compared to more than 500,000). On a closer look, however, this is understandable.

First, many of the errors occurring in the first experiment could be predicted and were easy to correct. For instance, the word *er* was always classified as a verb in the third (or first) person singular present indicative ('is, am'), as it usually is in Modern Icelandic. In Old Icelandic, however, this word is very often a temporal conjunction ('when') or a relative particle ('that, which'). When the tagger was trained on a corrected Old Icelandic text, it could quickly and easily learn the correct tagging of these words, due to their frequency.

Second, it is well known that tagging accuracy is usually very much lower for unknown words than for known words, and the number of unknown words was much lower in the second experiment. In the first experiment, using the model for Modern Icelandic, the unknown word rate was 14.6%, reflecting the fact that a number of Old Icelandic words are rare or do not occur in Modern Icelandic.

In the second experiment, using the model for Old Icelandic, the unknown word rate dropped to 9.6%, even though the training corpus was much smaller as pointed out above. This reflects the relatively small vocabulary of the Old Icelandic texts, which in turn reflects the narrow universe that the texts describe (cf. also Rögnvaldsson 1990).

Finally, we trained TnT on a union of the corrected Old Icelandic texts and the Modern Icelandic texts. Thus, the training set for the final experiment consists of around 500,000 words from Modern Icelandic texts plus 95,000 words from Old Icelandic texts. When we tagged the Old Icelandic corpus using this model, we got 92.7% accuracy for the same four samples as in the first two experiments.

Last year, Hrafn Loftsson (2013) took up the work on tagging Old Icelandic texts using tools developed for or trained on Modern Icelandic. He used the same training corpus as Rögnvaldsson and Helgadóttir (2011), although he improved the corpus by using some bootstrapping methods to correct a number of tagging errors. He tested two data-driven taggers, TriTagger and Stagger, one rule-based tagger, his own IceTagger which I will talk more about in a moment, and one hybrid tagger, comprising both TriTagger and IceTagger.

As might be expected, IceTagger, which obtained the second highest accuracy for Modern Icelandic, performed badly when tagging Old Icelandic. Loftsson (2013) points out that this was to be expected, because the hand-crafted rules of IceTagger have been developed to tag modern texts. In his further experiments, Loftsson thus only used the hybrid and data-driven taggers. When training on the corrected Old Icelandic corpus (95,000 words) and the Iceland-

ic Frequency Dictionary files, he reached 92.32% accuracy, using ten-fold cross-validation and voting.

When the website for the Tagged Icelandic Corpus was established, it was decided to have the Old Icelandic corpus accessible through that website – `mim.arnastofnun.is`. The texts were tagged in the same manner as the Modern Icelandic texts, using four different taggers – IceTagger, TriTagger, MXPOST, fnTBL – and the final tag selected by simple voting using CombiTagger. This work was carried out by Sigrún Helgadóttir. We have not measured the accuracy of the tagging.

7. Use of IceNLP in editing medieval texts

Hrafn Loftsson is the main author of the most important language technology tools that have been written for Modern Icelandic. He started developing an open source software package for analyzing and processing Icelandic texts during his Ph.D. studies from 2004-2007 (Loftsson 2007, 2008; Loftsson and Rögnvaldsson 2007). Since then, students at the University of Reykjavík and the University of Iceland have helped in developing individual components. The software, which goes by the name of IceNLP, is rule-based and uses heuristic methods which guess prepositional phrases and syntactic functions and use the acquired knowledge to force feature agreement where appropriate.

IceNLP is implemented in Java and consists of a tokenizer, an unknown word guesser, the part-of-speech tagger IceTagger, the lemmatizer Lemmald, the shallow parser IceParser, and a named-entity recognizer. Anton Karl Ingason is the main author of the lemmatizer (Ingason et al. 2008). Individual components of IceNLP can be run independently or the JAVA clusters in question connected directly to software that is being developed.

IceNLP can be used for various tasks, such as breaking up text into individual tokens, tagging each token with its morphosyntactic tag, finding the lemma of a particular word and returning a shallow phrase structure and labels indicating syntactic functions. The package is downloadable under the LGPL (GNU Lesser General Public License), either directly from SourceForge or through META-SHARE. It can also be used online at `nlp.cs.ru.is`.

IceNLP was written for Modern Icelandic texts and its dictionary assumes that words have Modern Icelandic spelling. However, a number of experiments with using IceNLP for analyzing Old Icelandic texts have already been carried out. Ludger Zeewaert (2014) at the Árni Magnússon Institute in Reykjavík has used IceNLP in preparatory work for a new edition of Njáls Saga. He will talk about this experiment in a talk at this workshop later today, so I won't say anything about it for now, except for mentioning that his experience of using IceNLP is positive.

Alex Speed Kjeldsen (2013) at the University of Copenhagen is preparing a new electronic edition of Icelandic original diplomas from around 1300 to 1450. The texts comprise around 70,000 running words. Kjeldsen starts by OCR scanning Stefán Karlsson's (1963) diplomatic edition. After proofreading and correcting the scanned text, it is normalized semi-automatically with the aid of a word list. In order to be able to use language technology tools developed for Modern Icelandic, the text is normalized to Modern Icelandic spelling.

The normalized text is then tagged and lemmatized with IceNLP. Kjeldsen has written a couple of simple scripts to cater for a number of systematic differences between old and Modern Icelandic, and to convert the IceNLP tagset to the tagset used by *Menota*. This is shown on the slide. In this example, both the lemmatization and the tagging of all the words is correct. Kjeldsen reports that the tagging accuracy is usually over 90%.

Finally, a special program will convert the normalized Modern Icelandic forms to normalized Old Icelandic forms. This is not as simple as converting old to Modern Icelandic, but with the aid of word lists, this can be done with high accuracy.

8. IcePaHC

In 2011, we released IcePaHC, a one million word parsed historical corpus of Icelandic (Wallenberg et al. 2011; Rögnvaldsson et al. 2012b). This corpus which is completely free and open contains fragments of 60 texts ranging from the late 12th century to the present. The stability of the morphology and the limited changes in the syntax are the reasons why it is both possible and feasible to build one treebank with texts spanning a period of ten centuries. If the morphological system had changed dramatically, it would have been difficult to apply the same annotation scheme to old and modern texts.

We decided to convert all our texts to Modern Icelandic spelling. There were two reasons for this. One was that this makes it possible to search for individual words without having to capture all possible spelling variants using fuzzy search, regular expressions and the like. The main reason was, however, that we wanted to use the IceNLP package for preprocessing. If we had given the package input in the original spelling of each text, the result of the preprocessing would have been much poorer.

All major texts from the medieval period have been published, although the editions are not always as good as one would wish. Many texts from the 16th up to the 19th century, however, have never been published. We decided in the beginning that we would only use texts from printed sources – it would have been prohibitively time- consuming and expensive to digitize texts from manuscripts.

A number of texts were in Modern Icelandic spelling and could be used as they were. However, the majority of them were either in standardized Old Norse spelling or diplomatic, and thus had to be changed. For the texts in the standardized Old Norse spelling, the task was rather easy, and a few simple scripts could be used to make most of the changes. The diplomatic editions were much harder. Some scripts and simple search-and-replace could help, but since the spelling in these texts is often highly irregular, we had to go over them word by word and correct them, which was rather tedious and time-consuming.

After running IceNLP we ran a few programs developed within the project to prepare the text for manual annotation. The PoS tagset was converted to a format nearly identical to the Penn Parsed Corpora of Historical English, the format of the labeled bracketing was converted to the Penn treebank format for compatibility with existing software and various structures were partially annotated using CorpusSearch revision queries (Randall 2005). Such partial annotation includes building the left edge of subordinate clauses whose right edge is subsequently determined by a human annotator.

Since we believe the model we used in the building of IcePaHC was highly successful, we wanted to extend this model to a related less-resourced language and build a Faroese Parsed Historical Corpus, FarPaHC (Rögnvaldsson et al. 2012b). Morphologically and syntactically, Faroese resembles Icelandic in many ways. This makes it more important to follow the guidelines for IcePaHC and saves a lot of time because this decreases the number of decisions that need to be taken in the annotation process.

Since the annotator in the FarPaHC project has experience from IcePaHC we do not have to spend time and money on training. If one can read Icelandic it is fairly easy to read Faroese as well so the language does not either slow us down. However, for semi-automatic parsing we

are not able to use parsers, morphological taggers and lemmatizers written for Faroese – we must rely on programs written for Icelandic, especially the IceNLP package which was used in the IcePaHC annotation. This decreases the parsing speed in the beginning of the project.

Nevertheless, we used IceNLP in our annotation process, even though the tagger and lemmatizer produced a number of errors that they would not on Icelandic data. To correct this effect as efficiently as possible, we specified a number of handwritten rules as we went along which slowed the process down at first. We focused on writing rules for the most common words, e.g., pronouns, quantifiers, auxiliaries, modal verbs and function words, such as complementizers and prepositions.

9. Conclusion

In this talk, I have briefly described a number of projects where language technology tools developed for or trained on Modern Icelandic have been successfully used in analyzing older stages of the language. A common feature of all these projects is that the texts being analyzed have been converted to Modern Icelandic spelling. This would hardly be a realistic option for many other languages, since the changes in word-forms and vocabulary will in most cases be too great.

In most cases, the researchers working on these projects have not had to do the conversion themselves – they could get hold of recent editions in Modern Icelandic orthography. It is thus very fortunate for us that such editions have recently become popular. However, conversion from texts in normalized Old Icelandic spelling is no big deal – it can be done automatically for the most part, using a few simple scripts.

If only facsimile or diplomatic editions are available, the conversion is a more challenging task, which necessarily takes some manual intervention at least. There are a number of approaches to this task, for instance the methods described by Bollmann (2012) and Bollmann et al. (2011). The approach used by Jón Friðrik Daðason, Kristín Bjarnadóttir and Kristján Rúnarsson (2014), which they will describe in their talk later today, is also very promising and will no doubt be very useful in such conversion.

Of course, a number of approaches have been followed to automatically enrich historical corpora with linguistic information and to use tools developed for modern languages to annotate older stages of the same languages. One approach is to use the existing tools without modification to annotate the unmodified historical texts, and then to correct the output manually. This is what has been done in the *Penn Historical Corpora*, for instance (Kroch and Taylor, 2000; Kroch Santorini and Delfs 2004).

Another approach is to normalize historical texts before tagging or parsing them with tools that have been developed for modern languages. The words are then projected to their modern equivalents prior to the annotation process and then eventually projected back to the original form afterwards. This is what has been done in a number of experiments with the IceNLP package (Rögnvaldsson et al. (2012b), Kjeldsen (2013), Zeevaert (2014)), and this is also what Scheible et al. (2011) have done in their work on Early Modern German, and Rayson et al. (2006) for Early Modern English.

The third approach is to start with tools developed for the modern language and retrain them on historical corpora using bootstrapping methods. In this case, the historical texts are also usually normalized to some extent. They can be normalized internally, so to speak, that is, normalized according to some artificial standard form for the stage of the language when they originated, as done by Dipper (2010; 2011) in work on Middle High German. This is also the

approach taken by Rögnvaldsson and Helgadóttir (2011) and Loftsson (2013) in their work on Old Icelandic texts described in this talk.

Sánchez-Marco, Boleda and Padro (2011) mention these three approaches and add the fourth one, which basically consists in expanding their Spanish dictionary with word variants from Old Spanish and retraining the tagger with a small training corpus. Hana, Feldman and Aharodnik (2011) perform a series of simple transformations to make a modern Czech text look more like a text in Old Czech and vice versa, and use a resource-light morphological analyzer to provide candidate tags. All kinds of mixtures of these methods are also possible.

As far as I can see, it is not possible to say which method is best. Different methods have been tested on different languages, the texts are different, the tools are different, the historical variation is different, etc., which makes a meaningful comparison difficult. Anyway, I will not venture to make such a comparison but I hope to have given a reasonably good overview of the experiments that have been made in applying language technology tools to Icelandic historical texts.

References

Benediktsson, H. (Ed.) (1972). *The First Grammatical Treatise.* Institute of Nordic Linguistics, Reykjavík.

Bjarnadóttir, K. (2012). The Database of Modern Icelandic Inflection. In *LREC 2012 Proceedings: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 – AfLaT 2012*.

Bollmann, M. (2012). (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, pp. 3–14. Lisbon, Portugal.

Bollmann, M., Petran, F. and Dipper, S. (2011). Rule-Based Normalization of Historical Texts. In *Proceedings of the RANLP Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pp. 34–42. Hissar, Bulgaria.

Daðason, J., Bjarnadóttir, K. and Rúnarsson, K. (2014). The Journal *Fjölnir* for Everyone: The Post-Processing of Historical OCR Texts. In *Proceedings of "Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage – LRT4HDA"*, workshop at the 9th International Conference on Language Resources and Evaluation, LREC 2014. Reykjavík.

Degnbol, H. (1985). Hvad en ordbog behøver – og andre ønsker [What a Dictionary Needs – and Others Wish for]. In *The Sixth International Saga Conference. Workshop Papers* I. Det arnamagnæanske institut, University of Copenhagen, Copenhagen, Denmark, pp. 235–254.

Dipper, S. (2010). POS-Tagging of Historical Language Data: First Experiments. In *Semantic Approaches in Natural Language Processing. Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10)*, pp. 117–121. Saarbrücken.

Dipper, S. (2011). Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison. In *Journal for Language Technology and Computational Linguistics, Special Issue, 26(2)*, pp. 25–37. (= Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities, 2012).

Eythórsson, T., Karlsson, B. and Sigurðardóttir, S.S. (2014). Greinir skáldskapar: A diachronic corpus of Icelandic poetic texts. In *Proceedings of "Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage – LRT4HDA"*, workshop at the 9th International Conference on Language Resources and Evaluation, LREC 2014. Reykjavík.

Faarlund, J.T. (2004). *The Syntax of Old Norse.* Oxford University Press, Oxford, UK.

Halldórsson, B., Torfason, J., Tómasson, S., Thorsson, Ö. (Eds.). (1985-86). *Íslendinga sögur* [The Icelandic Family Sagas]. Svart á hvítu, Reykjavík, Iceland.

Hana, J., Feldman, A. and Aharodnik, K. (2011). A Low-budget Tagger for Old Czech. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 10–18. Portland, OR.

Helgadóttir, S. (2005). Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2004*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 257–265.

Helgadóttir, S. (2007). Mörkun íslensks texta [Tagging Icelandic Text]. *Orð og tunga*, 9, pp. 75–107.

Helgadóttir, Sigrún, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir and Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). In: *Proceedings of "Language Technology for Normalization of Less-Resourced Languages"*, workshop at the 8th International Conference on Language Resources and Evaluation, LREC 2012, pp. 67–72. Istanbul, Turkey.

Helgadóttir, Sigrún, and Eiríkur Rögnvaldsson. 2013. Language Resources for Icelandic. In De Smedt, K., Borin, L., Lindén, K., Maegaard, B., Rögnvaldsson, E. and Vider, K. (Eds.): *Proceedings of the Workshop on Nordic Language Research Infrastructure at NODALIDA 2013*, pp. 60–76. NEALT Proceedings Series 20. Linköping Electronic Conference Proceedings, Linköping.

Ingason, A.K., Helgadóttir, S., Loftsson, H. and Rögnvaldsson, E. (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI), pp. 205–216. In A. Raante and B. Nordström (Eds.), *Advances in Natural Language Processing.* (Lecture Notes in Computer Science, Vol. 5221.) Springer, Berlin.

Karlsson, S. (Ed.) (1963). *Islandske originaldiplomer indtil 1450. Tekst*. [Icelandic Original Diplomas until 1450. Text.] Editiones Arnamagnæanæ 7, Series A. The Arnamagnæan Institute, Copenhagen.

Karlsson, S. (2004). *The Icelandic Language.* Viking Society for Northern Research, London.

Kjeldsen, A.S. (2013). Middelalderdiplomer – i en digital tid. En præsentation af et forskningsprojekt. [Medieval Diplomas – in the Digital Age. A Presentation of a Research Project.] MS, University of Copenhagen.

Kristjánsdóttir, B., Halldórsson, B., Sigurðsson, G., Grímsdóttir, G.Á., Ingólfsdóttir, G., Torfason, J., Tómasson, S., Thorsson, Ö. (Eds.). (1988). *Sturlunga saga* [The Sturlunga Collection]. Svart á hvítu, Reykjavík, Iceland.

Kristjánsdóttir, B., Halldórsson, B., Torfason, J., Thorsson , Ö. (Eds.). (1991). *Heimskringla* [The Sagas of the Kings of Norway]. Mál og menning, Reykjavík, Iceland.

Kristjánsdóttir, B., Rögnvaldsson, E. (editor), Ingólfsdóttir, G. and Thorsson, Ö. (1996). *Íslendinga sögur. Orðstöðulykill og texti.* [The Icelandic Family Sagas: Concordance and Text.] Mál og menning, Reykjavík, Iceland. [CD-ROM]

Kroch, A., Santorini, B., Delfs, L. (2004). Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html

Kroch, A., Taylor, A. (2000). Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html

Loftsson, H. (2007). Tagging and Parsing Icelandic Text. PhD thesis, Department of Computer Science, University of Sheffield.

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

Loftsson, H. (2013). Tagging the Past. Experiments Using the Saga Corpus. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA-2013), NEALT Proceedings Series 16*. Oslo.

Loftsson, H., Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In Nivre, J., Kaalep, H.-J., Muischnek, K. and Koit, M. (Eds.): *NODALIDA 2007 Conference Proceedings.* University of Tartu, pp. 128–135.

Pind, J. (Ed.), Magnússon, F., Briem, S. (1991). *Íslensk orðtíðnibók* [Icelandic Frequency Dictionary, IFD] Orðabók Háskólans, Reykjavík, Iceland.

Randall, B. (2005). CorpusSearch 2 Users Guide. University of Pennsylvania, Philadelphia. (*http:// corpussearch.sourceforge.net/CS-manual/Contents.html*).

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK.

Rögnvaldsson, E. (1990). Orðstöðulykill Íslendinga sagna [The Concordance to the Icelandic Family Sagas]. *Skáldskaparmál*, 1, pp. 54–61.

Rögnvaldsson, E. (1995). Old Icelandic: A Non-Configurational Language? *NOWELE*, 26, pp. 3-29.

Rögnvaldsson, E. (1996a). Word Order Variation in the VP in Old Icelandic. *Working Papers in Scandinavian Syntax*, 58, pp. 55–86.

Rögnvaldsson, E. (1996b). Frumlag og fall að fornu. [Subject and Case in Old Icelandic.] *Íslenskt mál*, 18, pp. 37–69.

Rögnvaldsson, E. 1997. Orðafar Íslendinga sagna. [The Vocabulary of the Icelandic Family Sagas.] In Agnarsdóttir, A., Pétursson, P. and Tulinius, T.H. (Eds.): *Milli himins og jarðar*, pp. 271-286. Háskólaútgáfan, Reykjavík.

Rögnvaldsson, E. (2000). Setningarstaða boðháttarsagna í fornu máli. [The Syntax of the Imperative in Old Icelandic.] *Íslenskt mál*, 22, pp. 63–90.

Rögnvaldsson, E. (2005). Setningafræðilegar breytingar í íslensku. [Syntactic Changes in Icelandic.] In Thráinsson, H. (Ed.) *Setningar. Handbók um setningafræði* [Sentences: A Handbook on Syntax]. (Íslensk tunga III.) Almenna bókafélagið, Reykjavík, Iceland, pp. 602–635.

Rögnvaldsson, E., Helgadóttir, S. (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In C. Sporleder, A.P.J van den

Bosch and K.A. Zervanou (Eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop series*. Berlin: Springer, pp. 63–76.

Rögnvaldsson, E., Jóhannsdóttir, K.M., Helgadóttir, S. and Steingrímsson, S. (2012a). *Íslensk tunga á stafrænni öld / The Icelandic Language in the Digital Age*. META-NET White Paper Series. Springer, Berlin.

Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J.C. (2012b). The Icelandic Parsed Historical Corpus (IcePaHC). *Proceedings of LREC 2012*, pp. 1978–1984. Istanbul, Turkey.

Sánchez-Marco, C., Boleda, G., and Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the* 5*th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR.

Scheible, S., Whitt, R. J., Durrell, M., and Bennett, P. (2011). Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *Proceedings of the* 5*th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR.

Wallenberg, J.C., Ingason, A.K., Sigurðsson, E.F. and Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC).* Version 0.9. http://www.linguist.is/icelandic_treebank

Zeewaert, L. (2014). IceTagging the "Golden Codex". Using language tools developed for Modern Icelandic on a corpus of Old Norse manuscripts. In *Proceedings of "Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage – LRT4HDA"*, workshop at the 9th International Conference on Language Resources and Evaluation, LREC 2014. Reykjavík.