The Icelandic Parsed Historical Corpus (IcePaHC)

Eiríkur Rögnvaldsson[1], Anton Karl Ingason[2], Einar Freyr Sigurðsson[1], Joel Wallenberg[3]

University of Iceland[1], University of Pennsylvania[2], Newcastle University[3]
Árnagarði við Suðurgötu, IS-101 Reykjavík[1], 619 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305[2], Percy Building, Newcastle University, Newcastle Upon Tyne, NE1 7RU[3]
E-mail: eirikur@hi.is, ingason@ling.upenn.edu, einarfs@gmail.com, joel.wallenberg@gmail.com

Abstract

We describe the background for and building of IcePaHC, a one million word parsed historical corpus of Icelandic which has just been finished. This corpus which is completely free and open contains fragments of 60 texts ranging from the late 12th century to the present. We describe the text selection and text collecting process and discuss the quality of the texts and their conversion to modern Icelandic spelling. We explain why we choose to use a phrase structure Penn style annotation scheme and briefly describe the syntactic annotation process. We also describe a spin-off project which is only in its beginning stages: a parsed historical corpus of Faroese. Finally, we advocate the importance of an open source policy as regards language resources.

Keywords: Icelandic, Faroese, treebank, parsed corpus, annotation

1. Introduction

The parsed corpus, or treebank, reported on in this paper, Icelandic Parsed Historical Corpus or IcePaHC (Wallenberg et al., 2011) is the product of three different projects which originally had different aims. The earliest and largest of these projects was a subpart of a large language technology project which had the aim of developing three different basic language resources for Icelandic. The aim of this subproject was to build a treebank of Modern Icelandic for use in language technology and to develop efficient parsing methods and tools for less resourced languages. Since some of the participants had been involved in historical syntax research, we also wanted to include a few texts from older stages of the language. However, the main emphasis was on language technology use – we intended to use the texts to train a statistical parser for Modern Icelandic.

At the same time, two other projects with the aim of developing resources for studying diachronic Icelandic syntax were in preparation. After some discussion, the participants in these three projects decided to join forces and make a combined effort to build a large parsed corpus covering the history of Icelandic syntax from the earliest sources up to the present. This corpus thus serves the dual purpose of being one of the cornerstones of Icelandic language technology and being an invaluable tool in Icelandic diachronic syntax research.

The corpus is now finished and contains one million words. It has been made available through free download (http://linguist.is/icelandic_treebank/Download) – in fact, we released preliminary versions every three months through the whole project period. We believe the corpus is unusual in many ways.

First, it is designed from the beginning to serve both as a language technology tool and a syntactic research tool, and developed by people with research experience in both diachronic syntax and computational linguistics. Most parsed corpora are developed either for language technology use (such as the Penn Treebank, http://www.cis.upenn.edu/~treebank/) or for syntactic research (such as the Penn Parsed Corpora of Historical English, PPCHE, http://www.ling.upenn.edu/hist-corpora/; Kroch and Taylor, 2000; Kroch, Santorini and Delfs, 2004).

Secondly, the corpus spans almost ten centuries – the oldest texts are written in the final decades of the 12th century and the youngest are from the first decade of the 21st century. As far as we know, no other single parsed corpus comes close to that. Most other languages have changed so much in the course of the last thousand years that it would be more challenging to apply the same annotation scheme for the whole period.

Third, our corpus contains over one million words and is thus among the largest parsed corpora that have been published for any language. As far as we know, only English and Czech have larger hand-checked treebanks.

Fourth, the corpus is completely free and open without any registration or paperwork, and the same goes for all the software that has been used to build it and the software that was developed within the project. Both the software and the corpus itself are distributed under the LGPL license.

This paper describes the background of the treebank. In the next section, we explain how it is possible and why it is feasible to build a diachronic treebank spanning almost ten centuries in the history of Icelandic. After that, we discuss several aspects of the material in the treebank – the selection of the texts, their quality, and their conversion to modern Icelandic spelling. We then go on to explain why we choose to build a Penn style treebank instead of a dependency treebank, which might perhaps seem a more obvious choice. Following a short description of the annotation process, we briefly describe a spin-off project: a parsed historical corpus of Faroese, FarPaHC, which is only in its beginning stages. Finally, we present our open source policy and set forth "10 basic types of user freedom" for language resources.

2. The diachronic dimension

Icelandic is a language with a rich literary heritage ranging from the 12th century to the present. 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. The phonetic quality of many of the vowels has changed, and the quantity system has changed such that vowel length is now context-dependent instead of being fixed.

On the other hand, the inflectional system and the morphology has in all relevant respects remained unchanged from the earliest texts up to the present, although a number of nouns have shifted inflectional class, a few strong verbs have become weak, one inflectional class of nouns has been lost, and the dual in personal and possessive pronoun has disappeared. The syntax is also basically the same, although a number of changes have occurred. The changes mainly involve word order, especially within the verb phrase, and the development of new modal constructions (cf. for instance Rögnvaldsson and Helgadóttir, 2011).

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 the same, it is usually relatively straightforward to convert older spelling to the modern standard and get legible text.

These two features – the stability of the morphology and the 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. On the other hand, the known syntactic changes and variation do not greatly complicate the annotation scheme, making it feasible to build a tool that enables us to study these changes and variation in a systematic way. The parsed historical corpus is such a tool.

3. Texts

3.1 Text selection

Selecting texts to parse for the corpus was a challenging task. We wanted to have the corpus both representative of different text genres and comparable through the centuries. This meant that we excluded some genres which have emerged only recently, such as newspaper texts. We decided in the beginning on a goal of parsing one million words – approximately 100,000 from each century of Icelandic literary tradition.

Our original plan was to have samples from five different genres of text for each century – preferably 20,000 words from each text. The genres we had in mind were narrative texts, religious texts, biographies, laws, and science. We knew from the beginning that it would be impossible to reach this goal, simply because texts belonging to some of the genres do not exist from all 10 centuries. We started with narrative texts and religious texts, since texts from these two genres were easiest to get hold of.

When we were well into the project, we decided to abandon the original plan and concentrate on these two genres. Narrative texts are the overwhelming majority of preserved medieval texts, and those which have been most studied and are easiest to get. It is also relatively easy to find religious texts from most centuries, but biographies, laws, and scientific texts are much fewer and harder to find in edited editions. Thus, we decided to stick to the original plan of having around 100,000 words from each century, but instead of dividing this evenly among five genres, we aimed at having 80,000 words of narrative texts and 20,000 words of religious prose. This also increases the data set for the two genres, allowing for more reliable studies of style-shifting phenomena.

By and large, this plan could be upheld. However, we didn't manage to find any religious text that could be attributed to the 15th century, and it proved to be difficult to find enough narrative texts from the 16th through 18th centuries. Instead, we included more of religious texts from the 16th century and some biographies from the 18th and 19th centuries. The distribution of the texts across genres and centuries is shown in table 1.

	nar	**rel**	**bio**	**sci**	**law**	Total
12th	0	40871	0	4439	0	**45310**
13th	93463	21196	0	0	6183	**120842**
14th	77370	21315	0	0	0	**98685**
15th	111560	0	0	0	0	**111560**
16th	35733	60464	0	0	0	**96197**
17th	46281	28134	52997	0	0	**127412**
18th	63322	22963	22099	0	0	**108384**
19th	100362	20370	0	3268	0	**124000**
20th	103921	21234	0	0	0	**125155**
21st	43102	0	0	0	0	**43102**
Total	**675114**	**236547**	**75096**	**7707**	**6183**	**1000647**

Table 1: Text types

The corpus contains (samples of) 60 different texts which came from various sources. Approximately 20 texts were taken from text repositories on the Internet, especially the Icelandic Netútgáfan (http://snerpa.is/net) but a few came from the Project Gutenberg website (http://www. gutenberg.org), the Internet Archive (http://www.archive.org/) and the Medieval Nordic Text Archive (http://www.menota.org/). Around 10 texts came from the Árni Magnússon Institute text archive (http://www.lexis.hi.is/corpus/). We received around 10 texts directly from scholars who have been editing them or publishing companies that had published them. The remaining texts,

around 20, were keyed in for us by students working on the project. Four texts from the 20th and 21st centuries are still under copyright, but we contacted the authors who gave us permission to use them.

3.2 Text quality

The quality of the texts varies a lot. Very few Old Icelandic texts are preserved in the original, and exact dating of the texts is often very difficult. Usually, the preserved manuscripts are assumed to be several decades and even centuries younger than the original text. We know that the scribes did not copy the manuscripts letter for letter – often they just used their own spelling instead of retaining the spelling of the original. This makes it very difficult to use the text to study phonology and morphology (cf. for instance Bernharðsson, 1999).

For those who use the text to study syntax and syntactic change, however, this is not a serious drawback, although in exceptional cases case distinctions in endings may be lost due to phonological changes and/or changes in spelling. On the other hand, it is usually assumed that scribes more or less retained the word order and other syntactic features of the manuscript they were copying, although there are a number of known exceptions to this. The treebank is accompanied by detailed "info" files which users can consult and make their own decisions on using or disregarding data from certain texts.

3.3 Text conversion

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 open-source IceNLP package for preprocessing. This package (available at http://icenlp.sourceforge.net*)* contains a tokenizer, a PoS tagger, a lemmatizer, and a shallow parser (Loftsson, 2008; Loftsson and Rögnvaldsson, 2007; Ingason et al., 2008). It was written for Modern Icelandic texts and its dictionary assumes that words have Modern Icelandic spelling. 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.

Editions of medieval Icelandic texts have one of three formats: 1) Diplomatic editions, where the text is printed exactly as in the manuscripts. 2) Standardized Old Norse spelling, which is a standard developed in the 19th century and is meant to mirror the sound system of 13th century Icelandic. 3) Modern Icelandic spelling. For most of the 20th century, editions of medieval texts intended for the public were usually in the standardized Old Norse spelling. Since the 1970s, however, it has become customary to use modern Icelandic spelling in new editions of medieval texts, even though editions mainly aimed at scholars usually try to mirror the spelling of the manuscript as closely as possible. Texts from the 19th century onwards usually only have minor deviations from the modern spelling.

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.

4. Annotation

4.1 Annotation scheme

One of the main questions which had to be answered before the annotation started was which annotation scheme to use. Most of the treebanks that have been built for the Scandinavian languages use some version of dependency parsing (e.g. Kromann, 2003; Bick, 2003; Nivre, Nilsson and Hall, 2006), so in some sense it would have been most natural for us to follow them. However, we had close contacts with the treebank team at the University of Pennsylvania from the early stages of the project, so it was a natural choice for us to use the phrase structure annotation scheme that they have developed for their parsed historical corpora (Kroch, Santorini and Delfs, 2004; Kroch and Taylor, 2000; Santorini, 2010). Thus, IcePaHC uses the same general type of labeled bracketing as the Penn Treebank (with dash-separated lemmata added) as shown below:

```
( (IP-MAT (NP-SBJ (PRO-N Hann-hann))
        (VBDI spurði-spyrja)
        (CP-QUE (WADVP-1 (WADV hvernig-hvernig))
              (C 0)
              (IP-SUB (ADVP *T*-1)
                    (NP-SBJ(NPR-D Grími-grímur))
                    (VBDS liði-líða))))
(ID 1888.GRIMUR.NAR-FIC,.301))
```

This proved to be a very fortunate decision. The Penn annotation scheme has already been adapted for Old English (Taylor et al,. 2003), which is rather similar to Icelandic in many respects, both as regards the syntax and the morphological system. Thus, the scheme could be applied to Icelandic with only slight modifications. Furthermore, the Penn team has written extensive annotation guidelines which were of tremendous help in our work (Santorini, 2010). We were careful to write our own guidelines and document all deviations from and additions to the Penn guidelines.

The decision to model our annotation on the Penn annotation system also meant that we could use the software that has been written especially to facilitate the annotation (CorpusDraw) and to search the corpus (CorpusSearch; Randall, 2005). An extra bonus is that it is now very easy to compare Icelandic and older stages of English. We can write search queries for English in CorpusSearch and by and large use the same queries for Icelandic, although minor modifications are sometimes necessary. Furthermore, Penn-style treebanks have been built or are currently being built for a number of other languages, such as French (Martineu et al., 2010), Portuguese (Galves and Britto, 2002), Early High German (Light, 2010), Classical Greek (Beck, 2011), Yiddish (Santorini, 1997/2008), and more. This development means that cross-linguistic, comparative diachronic studies can be carried out in a controlled and reproducible way with the same search queries across these datasets.

Yet another reason for choosing the Penn annotation system was that it is relatively rich, compared to most dependency-based schemes. Thus, it should – in principle, at least – be possible to convert our treebank to a dependency treebank, although some information will be lost in the conversion. Going the other way, that is, converting a dependency-based treebank to a Penn-style phrase structure treebank, would, on the other hand, be impossible without adding information.

Even though we followed the Penn scheme in most cases, we found it necessary to make some slight modifications, as mentioned above. The most important of these are that we annotate the words for lemma and nominals also for case, neither of which is done in the English historical corpora (excepting case in Old English).

4.2 The annotation process

After the texts had been converted to modern Icelandic spelling, they were handed over to student assistants who had the task of dividing them into clauses. Some periods do not signal the end of a tree and not all trees end in periods. Sentence boundary detection for English has been shown to classify periods such that 98.5% of sentences boundaries are correctly identified, a considerable improvement over the 90% precision of a baseline classifier which assumes every period to be a boundary (Palmer and Hearst, 1994). While this may seem encouraging we have two good reasons for preferring a manual approach to clause boundary detection.

First, while rules for classifying periods as boundary marking or not are fairly simple, the rules for inserting clause boundaries (usually between well-formed matrix clauses but sometimes sentence fragments without enough material to reconstruct clausal structure reliably and consistently) are more complicated and require interpretation of gapping structures where the nature and amount of omitted material affects the boundary status of conjunctive elements. Such problems can in principle be addressed using computational methods but the required tools are not currently available for Icelandic and their development was beyond the scope of our project.

Second, while clause boundary detection is not a trivial computational task it is fairly simple for a human and this part of the annotation could be carried out fast and reliably by research assistants which did not have to be trained in the complexities of full phrase structure annotation.

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.

The manual annotation phase comprised the bulk of the work. In the beginning, we used the CorpusDraw software to correct the parse, but we soon realized that it would be possible to speed up the annotation if we had software that was better suited for the task. Therefore, we wrote the annotation tool Annotald which made it possible to speed up the annotation considerably. Annotald is a browser based cross-platform visual tree editor which combines keyboard and mouse shortcuts such that the annotator can always keep the left hand on the keyboard and the right hand on the mouse. This avoids moving the right hand back and forth between mouse and keyboard which leads to improved speed and accuracy over CorpusDraw (see Figure 1 for the overall impact of improved methods and training on tree production). Annotald is released under the GPL license and continues to be developed by a growing team of programmers at the University of Pennsylvania (Beck, Ecay and Ingason, 2011).

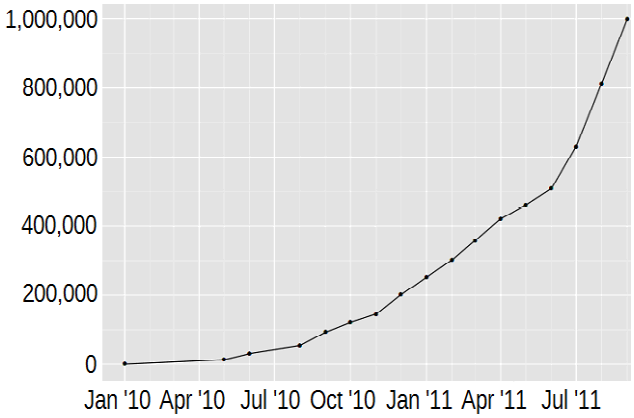

Figure 1: The annotation process

Three annotators worked on the project. In the beginning, they reviewed each other's work and spent a lot of time consulting the annotation manual for the Penn Historical Corpora (Santorini, 2010), which we succeeded in adapting to Icelandic. When the annotators had become well acquainted with the annotation scheme and developed special annotation rules for most of the cases where Icelandic deviates from Old(er) English, they stopped reviewing each other's annotations and placed the em-

phasis on speeding up the annotation as much as possible. Figure 1 shows the annotation progress for the whole project period.

We are in no doubt that the speed of the annotation process, and the fact that a large part of the annotation has not been reviewed, has resulted in a considerable number of annotation errors and discrepancies. The errors are nevertheless a small minority of potential errors. Our current approach to correction is to systematically enforce more constraints on well-formed structures. For example, the latest release of the corpus (Wallenberg et al. 2011) incorrectly contains 51 clauses with two phrases labeled as direct objects (NP-OB1). This is about 1% of the 4727 double object clauses in the corpus and in most of the cases one of the two objects should be labeled as an indirect object (NP-OB2).

These errors and many more have been corrected for the next release of the corpus. However, we want to emphasize that the corpus is meant to be used for quantitative research, not qualitative. It is not possible to take the parsing of any single sentence from the corpus and rely on it without reflection. The reasons are both that the text may be of disputable age and the parse may contain an error. However, we believe that the quantity of the text and its overall quality make the corpus safe to use in quantitative studies and the quality will only improve as future users notify us of errors they catch in their work.

5. A Spin-off: FarPaHC

5.1 Background

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. Faroese is spoken by 48,000 people in the Faroe Islands, 25,000 in Denmark and 5,000 people in Iceland (http://fo.wikipedia.org/wiki/Føroyskt_mál). The closest relative of Faroese is Icelandic and the two languages share a number of properties. Both languages have rich case morphology (four cases) and nominals and adjectives also inflect for number and gender. A native speaker of Icelandic can read written Faroese (and vice versa), though the spoken languages are just different enough to not be mutually intelligible without considerable effort.

Our goal is to build a 250,000 word fully parsed, PoS tagged and lemmatized corpus of 19th-21st century Faroese. The reason why we do not have older texts is that Faroese literary tradition is so short. Since IcePaHC consists mostly of religious texts and narratives, we focus on balancing FarPaHC for the same genres, which will help to control for social and stylistic factors when syntactic change and stability in Icelandic and Faroese are compared. This means that most texts in the treebank date from the 19th century to the present. In IcePaHC, around 100,000 words were parsed for each century with the exception that there is only a little more than 40,000 words from the 12th and the 21st century each. In light of this, our aim for FarPaHC is to parse 200,000 words in total from texts dating from the 19th and 20th century and 50,000 words from the 21st century. The main reason for our modest goal of 250,000 words is that there is only one annotator working on the project.

Our method of text selection within these genres is primarily in order to facilitate cross-linguistic comparison, especially with Icelandic, English, and other Germanic languages, but it is also partly opportunistic. We try to find texts that are already digitized because it costs money and time to key them in (and even more to transcribe manuscripts). We also focus on parsing texts that are parallel translations of texts found in IcePaHC and PPCHE, such as the Gospel of John and Acts of the Apostles from the New Testament. We are also going to parse at least one Faroese version of a narrative found in IcePaHC. Such a text could be one of the Icelandic family sagas, for instance The Saga of Grettir the Strong (*Grettis saga*). The size of the samples from each text is intended to be similar to those in PPCHE and IcePaHC, roughly 20,000 words (when possible).

5.2 Annotation

When we started building FarPaHC we used the modified PPCHE scheme from IcePaHC. As pointed out above, this allows for easy crosslinguistic comparison, as well as for experimentation in developing multilingual taggers and parsers. For more consistent parse we document our decisions (http://linguist.is/farpahc/) and where possible, we follow the guidelines written for PPCHE (http://www.ling.upenn.edu/hist-corpora/) and the IcePaHC documentation (http://linguist.is/icelandic_treebank/). This is also helpful for users of the treebank. 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 use IceNLP in our annotation process, even though the tagger and lemmatizer produce a number of errors that they would not on Icelandic data. To correct this effect as efficiently as possible, we specify a number of handwritten rules as we go along which slows the process down at first. We focus on writing rules for the most common words, e.g., pronouns, quantifiers, auxiliaries (HAVE and BE), modal verbs and function words, such as complementizers and prepositions.

When we have run IceNLP on the raw data, we take the output and manually parse it with Annotald (Beck, Ecay and Ingason, 2011; cf. above). The following is a fully

parsed (and handcorrected) token in FarPaHC (the sentence is from the Gospel of John). For convenience we leave out the glosses.

```
( (IP-MAT (CODE VS:I_1J)
          (PP  (P Í)
               (NP (ADJ-D fyrstu$) (D-D $ni)))
          (BEDI var)
          (NP-SBJ (N-N orð$) (D-N $ið))
          (. ,-,)))
```

Since FarPaHC is compatible with, e.g., IcePaHC and PPCHE the same, or very similar, queries can be used. To demonstrate this we show the same sentence from the Gospel of John as above in IcePaHC and the Penn Parsed Corpus of Early Modern English (Kroch, Santorini and Delfs, 2004), respectively.

```
( (IP-MAT (CODE VS:I_1J)
          (NP-SBJ-1 *exp*)
          (PP  (P Í)
               (NP (N-D upphafi)))
          (BEDI var)
          (NP-1 (N-N orð))
          (. ,-,))
  (ID 1540.NTJOHN.REL-BIB,184.2))

( (IP-MAT (PP  (P In)
               (NP (D the) (N beginnynge)))
          (BED was)
          (NP-SBJ (D the) (N worde))
  (. ,)) (ID TYNDNEW-E1-H,I,1J.6))
```

We plan to publish the first version, 0.1, on June 1st. Every four months after that we plan to release a new version, with more words, until we achieve our goal of 250,000 words. Even though the first release of the corpus will be rather small, probably around 50,000 words, it can be useful to have this resource in comparative syntactic research, especially since the first texts that we parse, the Gospel of John and the Acts of the Apostles, are found in other Penn style treebanks, such as IcePaHC and PPCHE. People who prefer dependency treebanks should be able to use it by converting our parse to a dependency parse. The first release will nevertheless probably be too small for computational linguists to train data-driven tools.

6. Maximizing distribution and user freedom

We believe strongly in the sharing of resources. True to that spirit, we decided at the beginning of the project that we wanted to make our work as open and widely distributed as possible. To emphasize that, we defined the following "10 basic types of user freedom":

1. Raw data available can be downloaded for local use (corpus not hidden behind a search interface)
2. Comprehensive documentation freely available online
3. Available without registration, user identification of some sort, or signing of contracts
4. Development process of corpus relies only on free/open source software tools (for transparent replication of annotation process)
5. Open development (annotation is carried out in an open online version control repository for transparency regarding the actual steps taken in the development and immediate access to work-in-progress)
6. Regular scheduled releases of numbered versions during development as well as for more permanent milestone versions so that researchers can always produce replicable results on a recent version of the corpus
7. Users can improve the corpus and release modified versions without special permission
8. Free of cost to academia
9. Free of cost to commercial users
10. Corpus released under a standard free license of some sort for straightforward compatibility with other projects (GPL, LGPL, CC, etc.)

We decided not to wait until the treebank was finished to release it. Instead, we released a new version every three months, in the hope of incrementally building up a user base and getting feedback from users which we could use to improve the treebank. This worked very well – for instance, Version 0.4, released in April 2011 and containing around 440,000 words, was downloaded (in different formats) more than 450 times from the project website (http://linguist.is/icelandic_treebank/Download). Furthermore, the treebank had already been used in a number of studies before the current version was released in August 2011 (e.g. Sapp, 2011; Ingason, Sigurðsson and Wallenberg, 2011; Light and Wallenberg, 2011).

As pointed out by Muhonen and Purtonen (2011), it is more difficult to gather information about the number of users of free treebanks than treebanks which require user registration, but we believe the benefits of user freedom outweigh such minor problems and in the course of time the impact of the treebank will be measured by its usage in published works rather than by raw download counts.

We follow the principles listed above strictly in the ongoing FarPaHC project. That involves regular scheduled releases before we reach our goal of 250,000 words. Regarding the fifth pointer above (open development), our open online version control repository is found on github. The URL is https://github.com/antonkarl/icecorpus.

Even though IcePaHC is practically finished, the current version is numbered 0.9 because some minor corrections remain to be made. The treebank is released in three versions; a zip-file containing the raw data of the of the corpus in labeled bracketing format; an easy-to-use setup executable for Windows that installs the corpus and a graphical user interface; and a zip-file containing the corpus and a platform independent user interface in Java.

The treebank has also been uploaded to the INESS repository at the University of Bergen (http://iness.uib.no) where it may be viewed and searched.

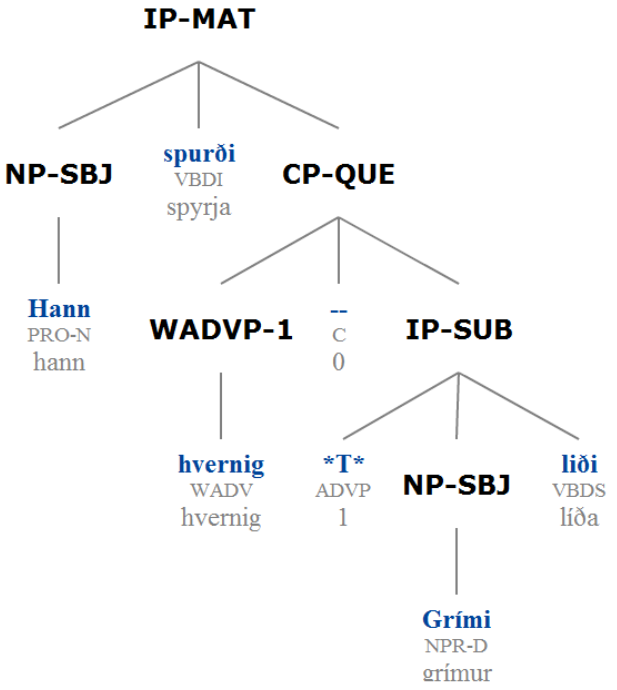

Figure 2: A sentence from IcePaHC in INESS

As shown in Figure 2, the sentences are presented in a familiar tree structure which is easier to read for most users than the output from CorpusSearch (see Sect. 4.1).

7. Conclusion

In this paper, we have described the parsed historical corpus of Icelandic, IcePaHC, and its motivations. As pointed out in the introduction, the corpus was built in order to serve two purposes: first, to be used within language technology to train parsers etc., and secondly, to be used as a tool for diachronic syntactic research.

Its usefulness for the latter purpose has already been demonstrated. For instance, four papers that were presented at the 13th Diachronic Generative Syntax Conference (DiGS) in June 2011 made use of the corpus (see http://www.ling.upenn.edu/Events/DIGS13/). As for the first purpose, the corpus has not yet been put to use but there is no reason to doubt that it can serve that purpose too. The corpus contains around 300,000 words which can safely be considered Modern Icelandic – texts from the 19th, 20th and 21st centuries. That is more than enough material to train a statistical parser. The project has also created a spin-off in the Faroese parsed historical corpus, FarPaHC.

As mentioned above, we believe that IcePaHC is unusual for a number of reasons. The most important one is that we have brought together a group of researchers who come from different fields and have different motives, but saw the benefits of joining their forces in building an important language resource which serves a dual purpose. The interdisciplinarity of the team should ensure that both humanist researchers and language technologists feel at ease in using the corpus in their work.

Finally, we emphasize the importance of distributing language resources under an open source license. This is especially important when working on less-resourced languages where duplication of work must be avoided. We hope that other researchers will follow in our steps and make their resources and tools publicly available for the benefit of all.

8. Acknowledgements

The building of IcePaHC was supported by the Icelandic Research Fund (Rannsóknasjóður), grant no 090662011, Viable Language Technology beyond English – Icelandic as a test case; the U.S. National Science Foundation (NSF) International Research Fellowship Program (IRFP), grant #OISE-0853114, Evolution of Language Systems: a comparative study of grammatical change in Icelandic and English; the University of Iceland Research Fund (Rannsóknasjóður Háskóla Íslands), grant Icelandic Diachronic Treebank (Sögulegur íslenskur trjábanki); and the EU ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement no 270899 (META-NORD). The Faroese Parsed Historical Corpus is funded by the University of Iceland Research Fund (Rannsóknasjóður Háskóla Íslands), grant Faroese treebank (Frumgerð færeysks trjábanka).

Thanks are due to several colleagues who generously gave us access to unpublished texts that they are editing. Thanks are also due to authors of copyrighted material who allowed us to use and distribute their texts. Thanks to Hrafn Loftsson who wrote most of the IceNLP software, to Brynhildur Stefánsdóttir and Hulda Óladóttir who assisted in parsing the texts, and to several students who keyed in a number of texts. Thanks to Victoria Rosén, Koenraad de Smedt and Paul Meurer at the University of Bergen for making IcePaHC a part of the INESS repository.

Thanks to anonymous reviewers for useful comments. Much of this material has previously been published in Rögnvaldsson et al. (2011), and IcePaHC has been presented at various occasions, such as the RILiVS workshop in Oslo in September 2009 (Rögnvaldsson, Ingason and Sigurðsson, 2011), talks at the University of Pennsylvania in Philadelphia, the University of Massachusetts at Amherst and New York University in May 2010, the annual conferences of the Institute of Humanities at the University of Iceland in Reykjavík in March 2011 and 2012, the MENOTA general assembly in Reykjavík in August 2011, the ACRH workshop in Heidelberg in January 2012, etc. We thank the audiences at these occasions for valuable discussion and comments.

Last but not least, we would like to thank our collaborators at the University of Pennsylvania, especially Tony Kroch and Beatrice Santorini, for their invaluable contributions to this work.

9. References

Beck, J.E. (2011). Penn Parsed Corpora of Historical Greek (PPCHiG). (http://www.ling.upenn.edu/~janabeck/greek-corpora.html).

Beck, J.E.; Ecay, A. and Ingason, A.K. (2011). Annotald, version 11.11. [Software for treebank annotation.] (http://github.com/janabeck/Annotald).

Bernharðsson, H. (1999). *Málblöndun í sautjándu aldar uppskriftum íslenskra miðaldahandrita.* Reykjavík: Institute of Linguistics, University of Iceland.

Bick, E. (2003). Arboretum, a Hybrid Treebank for Danish. In J. Nivre and E. Hinrichs (Eds.), *TLT 2003. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14-15 November 2003, Växjö, Sweden.* Växjö: Växjö University Press, pp. 9--20.

Galves, C., Britto, H.(2002). The Tycho Brahe Corpus of Historical Portuguese. Department of Linguistics, University of Campinas. Online publication, first edition. (http://www.tycho.iel.unicamp.br/~tycho/*)*.

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

Ingason, A.K.; Sigurðsson, E.F. and Wallenberg, J. (2011). Distinguishing Change and Stability: a Quantitative Study of Icelandic Oblique Subjects. Paper presented at DiGS 13, University of Pennsylvania, Philadelphia, June 3rd, 2011.

Kroch, A.; Santorini, B. and Delfs, L. (2004). The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, (http://www.ling.upenn.edu/hist-corpora/).

Kroch, A., Taylor, A. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, (http://www.ling.upenn.edu/hist-corpora/*)*.

Kromann, M.T. (2003). The Danish Dependency Treebank and the DTAG Treebank Tool. In J. Nivre and E. Hinrichs (Eds.), *TLT 2003. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14-15 November 2003, Växjö, Sweden.* Växjö: Växjö University Press, pp. 217--220.

Light, C. (2010). Parsed Corpus of Early New High German. (http://enhgcorpus.wikispaces.com/home).

Light, C., Wallenberg, J. (2011). On the Use of Passives across Germanic. Paper presented at DiGS 13, University of Pennsylvania, Philadelphia, June 4th, 2011.

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

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

Martineau, F.; Hirschbühler, P.; Kroch, A. and Morin, Y.C. (2010). Corpus MCVF (parsed corpus), Modéliser le changement: les voies du français, Département de français, University of Ottawa. CD-ROM, first edition (http://www.arts.uottawa.ca/voies/voies_fr.html).

Muhonen, K., Purtonen, T.K. (2011). Defining the Annotation Scheme of a Treebank: The End-Use Perspective. In Z. Vetulani (Ed.), *Human language technologies as a challenge for computer science and linguistics. Proceedings of the 5th Language and Technology Conference (LTC 2011), November 25-27, 2011.* Poznań, pp. 309--313.

Nivre, J.; Nilsson, J. and Hall, J. (2006). Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC).* Genoa, pp. 1392--1395.

Palmer, D.D., Hearst, M.A. (1994). Adaptive sentence boundary disambiguation. In *Proceedings of the fourth conference on Applied natural language processing.* Stroudsburg, PA: Association for Computational Linguistics, pp. 78--83.

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

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.; Ingason, A.K. and Sigurðsson, E.F. (2011). Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC). In J.B. Johannessen (Ed.), *Language Variation Infrastructure. Papers on selected projects*. Oslo Studies in Language 3.2. Oslo: University of Oslo, pp. 97--111.

Rögnvaldsson, E.; Ingason, A.K.; Sigurðsson, E.F and Wallenberg, J. (2011). Creating a Dual-Purpose Treebank. In Proceedings of the ACRH Workshop, Heidelberg, 5 Jan. 2012. *Journal for Language Technology and Computational Linguistics* 26(2), pp. 141--152.

Santorini, B. (1997/2008). The Penn Yiddish Corpus. University of Pennsylvania. For details, contact: beatrice@babel.ling.upenn.edu.

Santorini, B. (2010). Annotation manual for the Penn historical corpora and the PCEEC. University of Pennsylvania, Philadelphia. (http://www.ling.upenn.edu/hist-corpora/annotation/index.html).

Sapp, C. (2011). A Relative Pronoun in Old Norse? Paper presented at DiGS 13, University of Pennsylvania, Philadelphia, June 5th, 2011.

Taylor, A.; Warner, A.; Pintzuk, S. and Beths, F. (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York. (http://www.users.york.ac.uk/~lang22/YcoeHome1.htm).

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