Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, Joel Wallenberg

Creating a dual-purpose treebank

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. Furthermore, we advocate the importance of an open source policy as regards language resources.

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 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, *http://www.ling.upenn.edu/hist-corpora/*).

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 impractical to have text from such a long period in a single treebank.

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 brief description of the annotation process, we finally 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. The oldest preserved texts are mainly religious ones, such as for instance the Old Icelandic Homily Book (*Íslensk hómilíubók*, Stock. Perg. 4to no. 15), a large manuscript from around 1200, and a few translations from Latin.

In the 13th century, Icelanders started writing narrative texts, many of which are considered great literature and have been much celebrated. The most important of these texts are the Family Sagas (*Íslendingasögur*), stories about people living 300 years earlier, in the age of the settlement; *Heimskringla* (Sagas of the Kings of Norway) by the famous author Snorri Sturluson; *Sturlunga Saga*, a collection of stories about people and events in 13th century Iceland; and sagas of bishops.

The writing of these texts continued into the 14th century, but in the late 14th and 15th centuries, legendary sagas (*fornaldarsögur*) and romances (*riddarasögur*) become dominant, most of them translations from continental or English sources. However, Icelanders continued writing saga-style narratives on a small scale up to the 19th century. After the reformation in 1550, religious texts in the vernacular become more prominent. Some of the most important texts from the 17th and 18th centuries are biographies and travelogues. The first Icelandic novels were written in the first half of the 19th century.

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 and pointless 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 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, law, 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					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.

4 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 exceptions to this.

Most of the medieval texts that we used are taken from editions with a detailed introduction where the editor, usually a trained philologist, speculates about the dating of both the preserved text and the original. We have in most cases chosen to use the assumed dating of the original. If the scribes changed the syntax when copying older manuscripts – which they no doubt did occasionally – some syntactic features in some of the texts are actually younger and may thus lead us to believe that certain syntactic changes in fact occurred earlier than they actually did.

Another option would have been to use the dating of the preserved manuscript. That would have been misleading if it is assumed – as we do – that scribes usually didn't change the syntax when copying older manuscripts. Using the date of the manuscript, then, would lead us to believe that certain syntactic changes in fact occurred later than they actually did.

The third option would have been to use only those texts which can be dated fairly accurately and which exist in the original or in a manuscript close to the original in time. Unfortunately, this would have left us with only a couple of texts. None of the Sagas, for instance, is preserved in the original, and some of them only in manuscripts that are one, two, or even three centuries younger than the assumed writing of the text. Thus, we decided to choose quantity over quality in some cases and use texts which cannot be dated exactly and/or which only exist in manuscripts considerably younger than the original text. However, we always gave preference to the most reliably dated texts for a given time period when we had an option.

This may of course give rise to wrong or misleading results when the treebank is used to trace the origin or development of certain syntactic feature. However, 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. Of course, the treebank can also be used to check the dating of the texts. If, for instance, we are studying a certain syntactic phenomenon which increases or declines regularly through the centuries, but one text stands out as an exception, this might be an indication that this text has not been correctly dated.

These problems are by no means confined to our project – the developers of all historical treebanks are faced with similar problems. Note, however, that they have nothing to do with the construction and quality of the treebank per se. They only become problems when we want to interpret the results we get from searching the treebank for certain constructions and try to trace the development of a certain syntactic feature through the centuries. We are restricted by the available texts and have to interpret them somehow – we cannot ask for the native judgments of living speakers. This is of course one of the major problems that all diachronic syntacticians have to deal with.

5 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. Editions mainly aimed at scholars, however, 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.

6 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; EYTHÓRSSON and KARLSSON 2011), 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:

```
(1)( (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 (MARTINEAU 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).

7 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 reasons to prefer a manual approach. Failure to identify clause boundaries interacts with determining phrase structure as demonstrated by (2) below.

(2) a. We saw the problem that affected [NP the man and the woman] # and [NP the problem] made Jupiter look small.
b. We saw the problem that affected [NP the man] # and [NP the woman and the problem] made Jupiter look small.

A reviewer points out that it would be useful to quantify this problem but unfortunately we do not currently have reliable estimates. Nevertheless, we have two good reasons for our 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 pro-

duction). Annotald is released under the LGPL license and continues to be developed by a growing team of programmers at the University of Pennsylvania (the latest version is BECK et al. 2011).

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

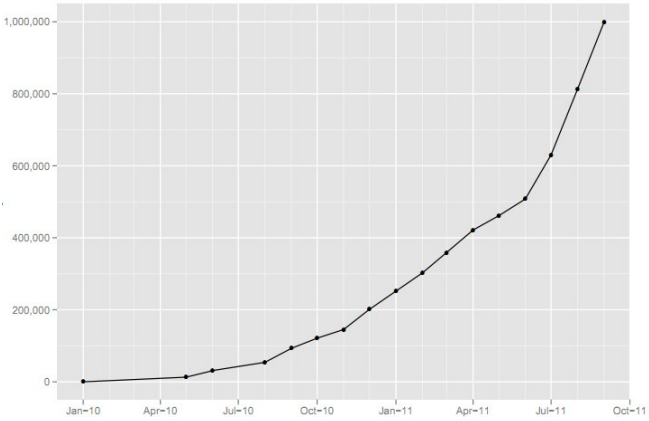

Figure 1: The annotation process

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.

8 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 (for instance SAPP 2011; INGASON, SIGURÐSSON and WALLENBERG 2011; LIGHT and WALLENBERG 2011).

Even though the treebank 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.

9 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.

As mentioned above, we believe that our corpus 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.

Litterature

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

BECK, J.E., A. ECAY and A.K. INGASON (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: Instite of Linguistics, University of Iceland.

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

EYTHÓRSSON, T., and B.M. KARLSSON (2011). Greinir skáldskapar: an Annotated Corpus of Old Icelandic Poetry. Paper presented „Bragarmál“, an international conference on Germanic and Icelandic metrics, University of Iceland, Reykjavík, September 24th, 2011.

GALVES, C., and H. BRITTO (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., S. HELGADÓTTIR, H. LOFTSSON and E. RÖGNVALDSSON (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In: Raante, A., and B. Nordström (eds.) (2008). Advances in Natural Language Processing. (Lecture Notes in Computer Science, Vol. 5221.) Berlin: Springer, 205-216.

INGASON, A.K., E.F. SIGURÐSSON and J. WALLENBERG (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., B. SANTORINI and L. DELFS (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., and A. TAYLOR (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: Nivre, J., and E. Hinrichs (eds.) (2003). TLT 2003. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14-15 November 2003, Växjö, Sweden. Växjö: Växjö University Press, 217-220.

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

LIGHT, C., and J. WALLENBERG (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):47-72.

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

MARTINEAU, F., P. HIRSCHBÜHLER, A. KROCH and Y.C. MORIN (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*).

NIVRE, J., J. NILSSON and J. HALL (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: 1392-1395.

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

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

RÖGNVALDSSON, E., and S. HELGADÓTTIR (2011). Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. In: Sporleder, C., A.P.J van den Bosch and K.A. Zervanou (eds.) (2011). Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop series. Berlin: Springer, 63-67.

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., A. WARNER, S. PINTZUK, AND F. BETHS (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., A.K. INGASON, E.F. SIGURÐSSON and E. RÖGNVALDSSON (2011). Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. (*http://www.linguist.is/icelandic_treebank*)