J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 97–112. (ISSN 1890-9639)
http://www.journals.uio.no/osla

COPING WITH VARIATION IN THE ICELANDIC PARSED HISTORICAL CORPUS (ICEPAHC)

EIRÍKUR RÖGNVALDSSON, ANTON KARL INGASON,
EINAR FREYR SIGURÐSSON
Reykjavík

ABSTRACT

We present an overview of an ongoing project which has the aim of developing methods for building a treebank of Icelandic. The treebank will contain texts from various different periods. Since Icelandic is an example of what has been called a less-resourced language when it comes to computational linguistics and language technology, it is essential to utilize the limited resources available as economically and efficiently as possible. We emphasize the importance of open source software and the interplay between linguistic knowledge and technological skills. We describe the workflow in the construction of the treebank and show how the different software tools work together towards the final representation. Finally, we show how the treebank can be used in studying some well known phenomena in Icelandic syntax.

[1] INTRODUCTION

In this paper, we describe an ongoing project, the Icelandic Parsed Historical Corpus (IcePaHC), which has the goal of developing economic and practical methods for building a treebank of Icelandic – methods which we hope can serve as a model in similar projects for other less-resourced languages. Icelandic is spoken by about 300,000 people and is clearly a less-resourced language (LRL) in any sense of the term. However, it has been the focus of much attention by syntacticians for the past two or three decades. There are several reasons for this. One is that due to its relatively rich morphology, Icelandic is ideal for testing several types of linguistic hypotheses. Another reason is that Icelandic has changed much less than its closest relatives and is thus ideal for testing and comparing theories of language change.

It is thus of great importance, not only to Icelandic syntacticians but to the general linguistic community, to have access to a large amount of well-structured data that enables researchers to study Icelandic syntax both synchronically and diachronically. As is well known, a syntactically parsed corpus – a treebank – is an important tool both for syntactic research and for the purposes of developing language technology tools. Our long-term goal is to build a treebank that will be useful both in syntactic research and for Icelandic language technology. The texts in our corpus will cover the history of Icelandic during a whole millennium, from the earliest written sources dating from the 12th century up to the present – approximately 100,000 words from each century.[1]

This paper describes our first steps in developing the treebanking methods and the treebank itself and is organized as follows. In section 2, we discuss briefly the motivations for building a parsed corpus like ours, and touch upon the challenges posed by the diversity of the texts. In section 3, we describe the software tools that we use and argue that an open source approach is essential for the development of NLP tools for less-resourced languages. In section 4, we describe the workflow in the construction of the treebank and show how the different software tools work together towards the final representation. Section 5 shows how the treebank can be used in the study of two well known phenomena in Icelandic syntax. Finally, section 6 is a conclusion.

[2] BACKGROUND AND CHALLENGES

Over the past two decades, interest in historical syntax has grown substantially among linguists. Accompanied by the growing amount of electronically available texts, this has led to the desire for – and possibility of – creating syntactically parsed corpora of historical texts, which could be used to facilitate search for examples of certain syntactic features and constructions. A few such corpora have been developed, the most notable being the Penn Parsed Corpora of Historical English, developed by Anthony Kroch and his associates (Kroch & Taylor 2000a; Kroch et al. 2004). These corpora have already proven their usefulness in a number of studies of older stages of English (cf., for instance, Kroch et al. 1995; Kroch & Taylor 2000b). We cooperate with the treebank team at the University of Pennsylvania and want to make our treebank compatible with their products – the Penn Treebank (Marcus et al. 1993) and the Penn Parsed Corpora of Historical English.

At a first glance, it may not seem feasible to build a diachronic treebank consisting of texts spanning a thousand years in the history of a language. However, Icelandic is often claimed to have undergone relatively small changes from the oldest written sources up to the present. The sound system, especially the vowel system, has changed dramatically, but these changes have not led to radical reduction or simplification of the system and hence they have not affected the

[1] At the time of the writing of this paper, a preview version (0.1) of the treebank (Wallenberg et al. 2010) which contains ca. 31,000 words from the 12th and 19th centuries has been released and can be downloaded from http://www.linguist.is/icelandic_treebank/Download. The corpus is released under the LGPL license which means that it may be freely distributed, modified and used in other software under certain restrictions - see http://www.gnu.org/licenses/lgpl.html. New versions will be released every three months until the project finishes, which is expected in mid-2011. We would like to thank Joel Wallenberg who has been instrumental in designing the corpus, Anthony Kroch and Beatrice Santorini for help and advice, and an anonymous reviewer for comments and suggestions.

inflectional system, which has not changed in any relevant respects. Thus, the morphosyntactic tagset developed for Modern Icelandic can be applied to earlier stages of the language without any modifications. 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 Icelandic vocabulary is still in use in Modern Icelandic, even though many words are confined to more formal styles and may have an archaic flavor.

On the other hand, many features of the syntax have changed (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. The diversity of the texts obviously poses quite a challenge to the project. It is clear that both the methods of construction, the annotation scheme, the query language, and the search software will have to be able to deal with considerable variation in sentence structure.

If the goal of a project is to construct a parsed corpus of a less-resourced language like Icelandic, it is important to utilize whatever resources are available as efficiently as possible. Actually, one could argue that all languages other than English are, to varying degrees, less-resourced with respect to English. Thus the problems that the less-resourced language faces with respect to language technology are shared among the languages of the world. One can identify two main kinds of problems for languages other than English:

- The amount of people and money available to develop resources is small compared to what is available for English.
- The language is different from English in important linguistic ways and thus the established state of the art solutions need to be adapted from how they are applied to English.

Ideally, we would have liked to put together a group of experts, each of which has substantial cross-disciplinary knowledge about parsed corpora, artificial intelligence / machine learning, generative syntax and perhaps some more. In reality, we have a few people who specialize in some of those fields. This means that one of the keys to successful treebank construction for a language like Icelandic is defining interfaces between the technological knowledge and the linguistic knowledge. We discuss our approach to this problem in section [3.2]. The importance of being able to employ linguistic knowledge relates to the typological difference between English and the LRL because Icelandic, for example, differs from English in ways that are relevant to parsing. Icelandic has a rich morphology that affects any kind of an annotation process as opposed to English where issues of morphology can be ignored to a great extent. Another example is the V2 (verb second) constraint that has a substantial effect on Icelandic word order. Such a constraint does not apply to Modern English.

Related to those problems is the fact that the small number of speakers of a language like Icelandic means that there is limited interest in the field from the commercial sector. In order to attract such interest some measures must be taken to make existing resources as accessible as possible. In our case the most important measures of this kind are the releasing of a complete language processing toolkit under a free and open source license, as discussed in section [3.1].

[3] TECHNOLOGY AND LINGUISTICS AS RESOURCES

[3.1] *Open Source BLARK as a Foundation*

It has been noted that in order to do any kind of work on language technology for a given language a set of some basic tools, referred to as a BLARK (Basic Language Resource Kit, (cf. Krauwer 2003)), is the minimum requirement. A few such tools have been developed for Icelandic and packaged under the name IceNLP. Those include a rule based PoS (Part-of-Speech) tagger (Loftsson 2008), an HMM (Hidden Markov Model) tagger, a shallow parser (Loftsson & Rögnvaldsson 2007), a lemmatizer (Ingason et al. 2008), a sentence segmentizer and a tokenizer. The IceNLP toolkit has recently been made open source (LGPL-licensed) to encourage further innovation in Icelandic language technology.[2]

Open source licenses are important for language technology in general as researchers have pointed out (e.g. Halácsy et al. 2007; Forcada 2006). We believe that this importance is even greater in the context of an LRL such as Icelandic. An accessible BLARK without serious licensing barriers can make a difference for the LRL in two important ways:

- It attracts researchers and commercial innovators to work on language technology for the LRL.
- It encourages linking the LRL with other international open source projects.

Many language technology projects focus on developing so-called language independent solutions for various tasks. Despite being language independent in nature those efforts are somewhat limited by the fact that practical aspects of setting up experiments for many languages always take time and therefore evaluation of the methods in question rarely extends to a large number of languages. We believe that a complete open source package of basic tools for a language like Icelandic makes the language much more feasible for inclusion in such experiments. A researcher can download IceNLP and start tagging and lemmatizing

[2] IceNLP can be downloaded from: http://sourceforge.net/projects/icenlp/

Icelandic text in minutes without having to consider licensing restrictions. Since IceNLP is LGPL-licensed it is also feasible for commercial software developers to include its features as part of their products. An open source license also encourages linking the BLARK of the LRL with other international open source projects and in the case of IceNLP there are already a few ongoing projects of this sort that would not have been possible without an open source BLARK.[3]

[3.2] *Automated Corpus Revision using Linguistic Terminology*

The IceNLP package includes a format conversion utility named Formald. One of the features of this utility is the ability to get a labeled bracketing representation of the output. Although such a conversion does not contribute anything to the information structure by itself it allows us to further manipulate the data using tools that are designed for working with labeled bracketing. One such tool that has been very useful in our annotation process is CorpusSearch (CS) (Randall 2005).

As the name implies CS is a tool that can be used to search parsed corpora but it can also be used for automated rule-based corpus revision. The main strength of CS in this respect is the fact that it is designed for linguists and the query language allows the user to interact with a treebank using terminology that is familiar to a syntactician. Relationships are expressed using terms like *dominates, precedes, c-commands*, etc. This means that CS provides an abstraction layer between linguistics and technology. A person who is trained in syntactic theory can develop an advanced rule-based parser without knowing much about the technology that does the computational work behind the scenes. In our case such a parser is built on top of the output of IceNLP. While such an abstraction layer may not be the most theoretically interesting fact about the annotation process of a treebank it means a great deal in terms of getting practical results with limited resources.

Automated corpus revision in CS is based on revision queries like the one shown in (1).

(1) **A CorpusSearch revision query**

```
query: ({1}[1]NP* hasSister {2}[2]NP-POS)
       AND ([1]NP* iPrecedes [2]NP-POS)

extend_span{1, 2}:
```

The query means: If any kind of an NP (NP*) has a sister in the tree that is an NP-POS and the first NP immediately precedes the latter, the span of the first NP

[3] Those include context sensitive spelling correction for Icelandic based on LanguageTool (Naber 2003), a machine translation system based on Apertium (Forcada 2006) and a commercial project that involves automated market research. Discussion of those projects is beyond the scope of this paper.

should be extended so that it includes the latter.

The syntax of regular search queries is the same as for revision queries except they do not include revision commands such as *extend_span*. The non-technical syntactician can therefore also use this syntax to search for suspicious patterns that probably need manual correction. For example one could construct a query that searches for IPs that include more than one subject, again using linguistic terminology.

[4] BUILDING THE TREEBANK

[4.1] *IceNLP*

The workflow we use in the construction of the diachronic parsed corpus of Icelandic makes extensive use of IceNLP as well as other open source software. To illustrate this let us take a look at how one sentence is processed using IceNLP. The sentence in (2) is an example from Old Icelandic.

(2) Rannveig og Hergerður voru dætur þeirra
Rannveig and Hergerður were daughters their
'Rannveig and Hergerður were their daughters'

The first step in the automated annotation is to run the sentence through IceTagger to assign PoS-tags as exemplified in (3).

(3) **Output from IceTagger:**

```
Rannveig nven-m
og c
Hergerður nven-m
voru sfg3fþ
dætur nvfn
þeirra fphfe
```

Since the IceNLP tools use the Icelandic tags (cf. Loftsson 2008) we keep this representation for now but the tags are translated into English in a later step.

In the second step we use IceParser to perform shallow parsing (chunking of phrases) as shown in (4). In addition to marking phrases IceParser annotates some syntactic functions such as subjects and objects.

(4) **Output from IceParser:**

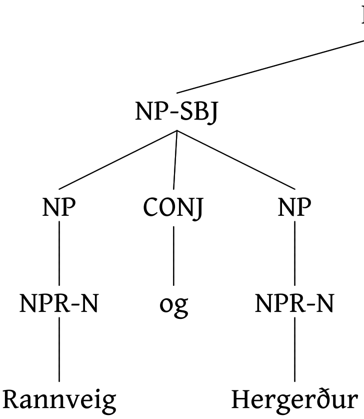
```
{*SUBJ> [NPs [NP Rannveig nven-m NP] [CP og c CP]
[NP Hergerður nven-m NP] NPs] *SUBJ>}
[VPb voru sfg3fþ VPB] {*COMP< [NP dætur nvfn NP] *COMP<}
{*QUAL [NP þeirra fphfe NP] *QUAL} . .
```

Then we use the lemmatizer Lemmald to assign a base form to each token in the sentence. The final step involving the IceNLP toolkit is to use one of its format conversion features to get a labeled bracketing representation of the sentence and translate the tagset to an annotation scheme that is mostly compatible with the Penn Corpora of Historical English. The result of those operations can be seen as labeled bracketing in (5) and as a tree diagram in (6). Note that lemmas are omitted from the tree diagrams in this paper.

(5) **Output after lemmatization and conversion to labeled bracketing:**

```
( (IP-MAT (NP-SBJ (NP (NPR-N Rannveig-rannveig) )
(CONJ og-og) (NP (NPR-N Hergerður-hergerður) ) )
(VPB (BEDI voru-vera) )
(NP-PRD (NS-N dætur-dóttir) )
(NP-POS (PRO-G þeirra-það) ) (. .-.) ) )
```

(6)



Thus, the diagram in (6) represents the kind of structure we can annotate using only the tools of the IceNLP toolkit.

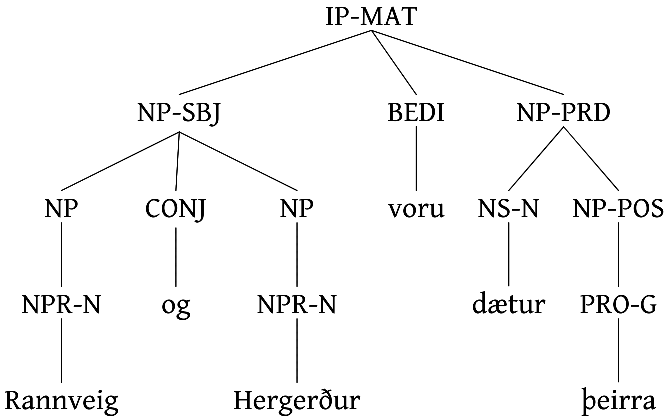
[4.2] *CorpusSearch and CorpusDraw*

The structure already contains a lot of information about the sentence but in order to finish the tree we use CS to apply revision queries to the structure. First we run the query in (1) so that NP-POS is moved under the immediately preceding NP. Finally we want the finite verb to be the head of the IP so we run the revision in (7) to delete VPs that are dominated by IPs and dominate finite verbs. Note that **finiteVerb** is defined by a regular expression that matches all finite verbs.

(7) **Revision query that removes extra VPs**

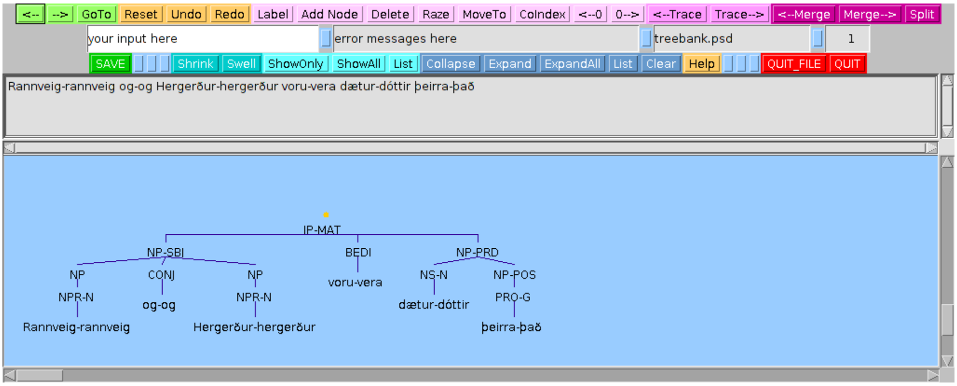
```
query: (IP-MAT iDoms {1}[1]VP*)
        AND ([1]VP* iDoms finiteVerb)

delete_node{1}:
```

(8)



The resulting tree is shown in (8). In this case, the automated rule-based annotation manages to generate the full structure we want. This is of course not always the case and while we aim to cover as many types of structure as possible automatically there are various examples of incomplete or wrong annotation that require manual corrections. Again, we find it important to design the workflow in a way that does not require a lot of technical knowledge, especially the parts that require extensive understanding of the theory of syntax (manual corrections occur at the linguistic end of the abstraction layer, not the technical one!). For this we use CorpusDraw, a program that is bundled with CS and provides a visual interface for correcting trees. A screenshot of the previous sentence from CorpusDraw is shown in (9).

(9) **CorpusDraw screenshot**


<-- --> GoTo Reset Undo Redo Label Add Node Delete Raze MoveTo CoIndex <--0 0--> <--Trace Trace--> <--Merge Merge--> Split
your input here
error messages here
treebank.psd
1
SAVE Shrink Swell ShowOnly ShowAll List Collapse Expand ExpandAll List Clear Help QUIT_FILE QUIT
IP-MAT
NP-SBJ
BEDI
NP-PRD
NP
CONJ
NP
NS-N
NP-POS
NPR-N
og-og
NPR-N
voru-vera
dætur-dóttir
PRO-G
Rannveig-rannveig
Hergerður-hergerður
þeirra-það

Rannveig-rannveig og-og Hergerður-hergerður voru-vera dætur-dóttir þeirra-það

[5] TWO CASE STUDIES

In this section, we show how the treebank could be used to study two phenomena in Icelandic syntax – DAT-NOM verbs and the so-called New Passive.

[5.1] *DAT-NOM verbs*

In Modern Icelandic (MIce) we get variation between number agreement and non-agreement with DAT-NOM verbs that take plural nominative objects, cf. (10-a) and (10-b).

(10) a. Stelpunni líku**ðu** strákar
girl-THE-DAT liked-3-**PLUR** boys-NOM
'The girl liked boys'
b. Stelpunni líka**ði** strákar
girl-THE-DAT liked-3-**SING** boys-NOM

This kind of variation is also found in Old Icelandic (OIce) (Eythórsson & Jónsson 2009), although nominative agreement seems to have been more frequent in OIce. Non-agreement, on the other hand, seems to be much more frequent in MIce than in OIce. That would indicate a change over the ages – a change we can study in IcePaHC. If there has been a significant change in the agreement system, that might explain why, for some speakers, DAT-ACC (for original DAT-NOM verbs, e.g. *líka* 'like') seems to be grammatical (Árnadóttir & Sigurðsson 2008).

We do a CorpusSearch query to see if a change has taken place. To get a clear idea of what we are dealing with, let us first look at the raw data we get from CorpusDraw for (10-a) above:

(11) **Raw data**

```
( (IP-MAT (NP-SBJ (PRO-D Stelpu$-stelpa) (D-D $nni-hinn))
          (VBDI líkuðu-líka)
          (NP-OB1 (NS-N strákar-strákur))
          (. .-.)))
```

Now we can define the search which finds agreement as well as non-agreement with plural objects of DAT-NOM verbs, cf. (12).

(12) **A CorpusSearch query for DAT-NOM verbs**

```
node: IP*
 query: (IP-MAT*|IP-SUB* iDoms NP-SBJ)
        AND (IP-MAT*|IP-SUB* iDoms NP-OB1)
        AND (IP-MAT*|IP-SUB* iDoms !VAN*)
        AND (NP-SBJ iDoms *-D)
        AND (NP-OB1 iDoms NS-N)
```

The query matches any main clause (IP-MAT*) or embedded clause (IP-SUB*) that immediately dominates (iDoms) a subject (NP-SBJ) and an object (NP-OB1) and which does not immediately dominate a passive participle (!VAN*) (the '!' negates the matched element) since we do not want to include passives in our results. Furthermore, the subject phrase immediately dominates a nominal element in the dative case (*-D), where the star is a wildcard that matches nouns, pronouns, determiners and quantifiers. The object phrase immediately dominates a plural nominative noun (NS-N).

After running the CS query, we are able to compare relative frequencies of agreement vs. non-agreement from different periods of the history of Icelandic.

Since the treebank is compatible with the Penn Parsed Corpora of Historical English same, or similar, phenomena in Icelandic and English at various stages can be compared. Let us, for example, take a look at the raw data for the following sentence from Early Modern English (Kroch et al. 2004):

(13) **Early Modern English raw data**

```
( (IP-MAT (NP-SBJ (PRO I))
          (VBP believe)
          (CP-THT (C 0)
                  (IP-SUB (NP-SBJ (PRO I))
                  (MD shall)
                  (VB like)
                  (NP-OB1 (PRO$ your) (N cook))
                  (ADVP (ADV very) (ADV well))))
          (. .)) (ID FHATTON-E3-H,I,148.34))
```

In this example we have the main verb *like* in the embedded clause (IP-SUB). As can be seen, the parsing and the labels are (almost) the same as in IcePaHC. That makes it a lot easier to do a comparative study of those languages.

[5.2] *The New Passive*

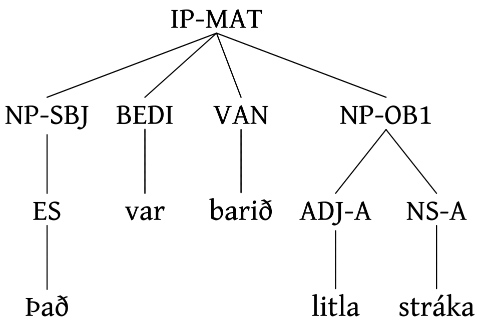
There has been a lively discussion about the New Passive in Icelandic in the recent years and opinions differ widely on its nature. Some researchers claim it is not a passive at all, but instead an active (cf. Maling & Sigurjónsdottir 2002), whereas others claim that it is simply a new form of the passive (e.g. Eythórsson 2008). Although it uses passive morphology, the object always stays in situ and does not undergo NP-movement (A-movement), cf. (15). Furthermore, it can be in the accusative case, and does not trigger agreement of the finite verb and the participle as a nominative argument in the canonical passive would do. Instead, the finite verb is always 3sg and the passive participle (the main verb), which assigns case (contra Burzio’s Generalization), is neuter singular. (14) shows different versions of the canonical passive, whereas (15) shows the New Passive.

(14) a. Það voru barðir litlir
it were beaten-MASC-PLUR little-MASC-NOM-PLUR
strákar
boys-MASC-NOM
‘Little boys were beaten’
b. Það voru litlir strákar barðir
c. Litlir strákar voru barðir

(15) a. Það var barið litla stráka
it was beaten-NEUT-SING little-MASC-ACC-PLUR boys-MASC-ACC
‘Little boys were beaten’
b. *Það var lítinn strák barið
it was little-MASC-ACC-SING boy-MASC-ACC beaten-NEUT-SING
‘A little boy was beaten’ (Eythórsson 2008, 213, ex. (76))

In the tree diagram in (16) we show how (15-a) would be parsed in our corpus. Notice that the expletive *það* is tagged ES.

(16)



As in the canonical passive, the verb *vera* 'be' (or *verða* 'will be, become') is always used in the New Passive. As expected, it is not always finite.

(17) Það hefur oft verið barið litla
it has often been beaten-NEUT-SING little-MASC-ACC-PLUR
stráka
boys-MASC-ACC
'Little boys have often been beaten'

Even though the New Passive sentences begin with the expletive *það* in the examples above, this is not always so, as seen in (18) and (19). Furthermore, as shown in (19), the passive participle does not always follow *vera/verða*. It can precede the verb in sentences where Stylistic Fronting has applied. In the absence of an overt expletive *það* we include an empty category **exp** in the annotation.

(18) Í gær var barið litla
yesterday was beaten-NEUT-SING little-MASC-ACC-PLUR
stráka
boys-MASC-ACC
'Little boys were beaten yesterday'

(19) Skoðað verður miða við innganginn
inspected-NEUT-SING will.be tickets-MASC-ACC on entrance-THE
'Tickets will be inspected on entering' (Maling 2006, 200, ex. (7))

Example (19) above shows that we cannot rely on the main verb immediately preceding the object.

From the facts described above we can use the following for a New Passive search query (with accusative object):

(20) a. It contains an expletive (overt or covert)
b. It contains the verb *vera* 'be' (BE*) or *verða* 'will be, become' (RD*)
c. It contains a passive participle (tagged as VAN)
d. It contains an object (NP-OB1)
e. The direct object is in accusative case

Following these facts (in that order), the CS query might look like this:

(21) **A CorpusSearch query for the New Passive**

```
node: IP*
query: (IP* idoms NP-SBJ)
       AND (NP-SBJ idoms ES|\*exp\*)
       AND (IP* iDoms BE*|RD*)
       AND (IP* iDoms VAN)
       AND (VAN hasSister NP-OB1)
       AND (NP-OB1 iDoms *-A)
```

Even though the literature on the innovative New Passive – which is almost exclusively found in texts from the late 20th century up to the present day – is already quite extensive, many things regarding its nature and origin remain unclear and disputed. The question arises, of course, why a 20th century child would re-analyse passive sentences. In other words, what is the source of the New Passive? Various attempts – which will not be repeated here – have been made to answer the question.

One possible factor could be that the use of the expletive *það* increased heavily in the early 19th century as Hróarsdóttir (1998) shows (cf. also Rögnvaldsson 2002). This can lead to the subject of the canonical passive not being A-moved, as shown in (14a). These possible effects on the New Passive cannot be fully investigated without a diachronic treebank.

[6] CONCLUSION

In this paper we have presented the outlines of our work in developing efficient methods for building a treebank of a less resourced language – Icelandic in our case. This is still very much a work in progress but we think that our approach could serve as an example for other less-resourced languages. We have emphasized the re-use of existing tools and the importance of open source policy in this respect. We have also emphasized the importance of linguistic insights and the interplay between linguistic knowledge and technological skills in developing software tools for building syntactic trees. We described the workflow in the construction of IcePaHC and presented examples of how it can be used to study celebrated constructions in Icelandic.

Obviously, we are far from having a full-fledged treebank at our disposal. However, we feel that we have come quite far in developing the methods for building the treebank, and we have already started the actual production of trees.

REFERENCES

Árnadóttir, H. & E. F. Sigurðsson. 2008. The glory of non-agreement: The rise of a new passive. Ms., University of Iceland.

Eythórsson, T. 2008. The New Passive in Icelandic really is a passive. In T. Eythórsson (ed.), *Grammatical Change and Linguistic Theory. The Rosendal papers*, 173–219. Amsterdam: John Benjamins.

Eythórsson, T. & J. G. Jónsson. 2009. Variation in Icelandic morphosyntax. In A. Dufter, J. Fleischer & G. Seiler (eds.), *Describing and Modeling Variation in Grammar*, Trends in Linguistics. Studies and Monographs 204, 81–96. Berlin: Mouton de Gruyter.

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

Forcada, M. L. 2006. Open-Source Machine Translation: an Opportunity for Minor Languages. In *Strategies for Developing Machine Translation for Minority Languages (5th SALTMIL Workshop on Minority Languages) (organized in conjunction with LREC 2006)*, 8–15. Genova.

Halácsy, P., A. Kornai & C. Oravecz. 2007. Hunpos – an open source trigram tagger. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, 209–212.

Hróarsdóttir, T. 1998. *Setningafræðilegar breytingar á 19. öld. Þróun þriggja málbreytinga*. Reykjavík: Institute of Linguistics, University of Iceland.

Ingason, A. K., S. Helgadóttir, H. Loftsson & E. Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *GoTAL '08: Proceedings of the 6th international conference on Advances in Natural Language Processing*, 205–216. Berlin, Heidelberg: Springer-Verlag. doi: http://dx.doi.org/10.1007/978-3-540-85287-2\\\\\\\_20.

Krauwer, S. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, 8–15.

Kroch, A., B. Santorini & L. Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition. Size: 1.8 million words.

Kroch, A. & A. Taylor. 2000a. Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition. Size: 1.3 million words.

Kroch, A. & A. Taylor. 2000b. Verb-Object Order in Early Middle English. In *Diachronic Syntax: Models and Mechanisms*, 132–163. Oxford University Press.

Kroch, A. S., A. Taylor & D. Ringe. 1995. The Middle English verb-second constraint: a case study in language contact and language change. In Susan Herring et al (ed.), *Textual Parameters in Older Language*. Amsterdam: John Benjamins.

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

Loftsson, H. & E. Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the* 16^{th} *Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*. Tartu, Estonia.

Maling, J. 2006. From passive to active. Syntactic change in progress in Icelandic. In B. Lyngfelt & T. Solstad (eds.), *Demoting the Agent. Passive, middle and other voice phenomena*, 197–223. Amsterdam: John Benjamins.

Maling, J. & S. Sigurjónsdóttir. 2002. The 'new impersonal' construction in Icelandic. *Journal of Comparative Germanic Linguistics* 5. 97–142.

Marcus, M. P., B. Santorini & M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330.

Naber, D. 2003. *A Rule-Based Style and Grammar Checker*. Diploma thesis, University of Bielefeld. URL http://www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf.

Randall, B. 2005. *Corpussearch 2 user's guide*. University of Pennsylvania. URL http://corpussearch.sourceforge.net/CS-manual/Contents.html.

Rögnvaldsson, E. 2002. ÞAÐ í fornu máli – og síðar. *Íslenskt mál* 24. 7–30.

Rögnvaldsson, E. 2005. Setningafræðilegar breytingar í íslensku. In H. Thráinsson (ed.), *Setningar. Handbók um setningafræði. Íslensk tunga* 3, 602–635. Reykjavík: Almenna bókafélagið.

Wallenberg, J. C., A. K. Ingason, E. F. Sigurðsson & E. Rögnvaldsson. 2010. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.1. Size: 31 thousand words. URL http://www.linguist.is/icelandic_treebank.

AUTHOR CONTACT INFORMATION

Eiríkur Rögnvaldsson
University of Iceland
Dept. of Icelandic
IS-101 Reykjavík
Iceland
eirikur@hi.is

Anton Karl Ingason
University of Iceland
Dept. of Icelandic
IS-101 Reykjavík
Iceland
anton.karl.ingason@gmail.com

Einar Freyr Sigurðsson
University of Iceland
Dept. of Icelandic
IS-101 Reykjavík
Iceland
einarfs@gmail.com