A Universal Dependencies Conversion Pipeline for a Penn-format Constituency Treebank

Þórunn Arnardóttir **Hinrik Hafsteinsson** **Einar Freyr Sigurðsson**
University of Iceland The Árni Magnússon Institute for Icelandic Studies

Kristín Bjarnadóttir
The Árni Magnússon Institute for Icelandic Studies

Anton Karl Ingason
University of Iceland

Hildur Jónsdóttir **Steinþór Steingrímsson**
The Árni Magnússon Institute for Icelandic Studies

Abstract

The topic of this paper is a rule-based pipeline for converting constituency treebanks based on the Penn Treebank format to Universal Dependencies (UD). We describe an Icelandic constituency treebank, its annotation scheme and the UD scheme. The conversion is discussed, the methods used to deliver a fully automated UD corpus and complications involved. To show its applicability to corpora in different languages, we extend the pipeline and convert a Faroese constituency treebank to a UD corpus. The result is an open-source conversion tool, published under an Apache 2.0 license, applicable to a Penn-style treebank for conversion to a UD corpus, along with the two new UD corpora.

1 Introduction

The Universal Dependencies (UD) project (Nivre et al., 2016) in version 2.6 consists of 163 treebanks in 92 languages and its standardized annotation scheme makes it an appealing option when creating a treebank. Different methods are available to create a UD treebank, ranging from manual parsing to manual correction of automatic parsing. A UD treebank can also be automatically converted from an existing treebank, which uses a different annotation scheme, as in the present work. We describe a conversion pipeline for a constituency treebank, the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al., 2011; Rögnvaldsson et al., 2011, 2012). We also apply the pipeline to a smaller corpus, the Faroese Parsed Historical Corpus (FarPaHC; Ingason et al., 2012; Ingason et al., 2014), which contains texts in another language but uses the same annotation scheme.[1] With minimal modifications to the pipeline, the conversion is unaffected by linguistic differences in the corpora, which underlines the possibility of using it for other corpora, which use the same scheme.

The conversion uses NLTK (Bird et al., 2009) as a base and it is specific to the annotation scheme of IcePaHC, which itself is based on the Penn Parsed Corpora of Historical English (PPCHE; Kroch and Taylor, 2000; Kroch et al., 2004). Some conversion tools are already available for converting a constituency treebank to the UD scheme, e.g. the PyStanfordDependencies tool[2] and the LTH Constituent-to-Dependency Conversion Tool for Penn-style Treebanks (Johansson and Nugues, 2007).[3] Both tools convert a treebank in the Penn Treebank format to a dependency format. The annotation scheme of the PPCHE is based on the Penn Treebank scheme but the two are not similar enough to be used unaltered with the pre-existing software, so the conversion between the two schemes was deemed to be impractical. Therefore, our conversion tool takes as input an unaltered IcePaHC file and delivers its UD-annotated counterpart. Both IcePaHC and FarPaHC are open-source, published under a CC BY 4.0 license, and we maintain that policy, publishing the conversion tool under an Apache 2.0 license and the resulting UD corpora under a CC BY-SA 4.0 license. The tool is reusable for corpora annotated in the same manner as the constituency treebanks and the source code is available on GitHub.[4] The resulting corpora will be

[1] Note that an Icelandic UD treebank (Jónsdóttir and Ingason, 2020) and a Faroese one (Tyers et al., 2018) already exist.

[2] https://github.com/dmcc/PyStanfordDependencies

[3] http://nlp.cs.lth.se/software/treebank_converter/

[4] https://github.com/thorunna/UDConverter

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 16–25
©2020 Association for Computational Linguistics

included in the next version of UD, version 2.7, to be released on November 15, 2020.

The paper is structured as follows. Section 2 describes IcePaHC and its annotation scheme and Section 3 compares PPCHE and UD. Section 4 describes the conversion itself, and its stages, i.e. how the files are prepared for the converter (4.1), how the PoS tags and morphological features are mapped (4.2), and how the relational information of heads and dependency are extracted from the IcePaHC format (4.3). Section 5 discusses the FarPaHC conversion, Section 6 considers further use of the converter and the UD treebanks and Section 7 concludes.

2 The Icelandic Parsed Historical Corpus

IcePaHC is a one-million-word, diachronic corpus, which includes texts from the 12th to 21st centuries (Rögnvaldsson et al., 2012). The trees in the corpus have been manually corrected according to the PPCHE annotation scheme, which uses labeled bracketing in the same way as the Penn Treebank. Some minor adjustments were made to adapt the annotation scheme to Icelandic grammar. The PoS tagset used is based on the one particular to the PPCHE, with minor changes to adapt it to Icelandic grammar. The manual IcePaHC annotation process is built on a number of automatic pre-processing steps involving available Natural Language Processing tools for Icelandic, including PoS tagging and shallow parsing in IceNLP (Loftsson and Rögnvaldsson, 2007), and the lemmatizer Lemmald (Ingason et al., 2008).

The IcePaHC scheme splits sentences into matrix clauses and marks their phrases. The phrases and their tokens are marked according to the tagset, which consists of 43 tags and their function tags.[5] Moreover, every token's lemma is displayed along with traces and empty phrases. Figure 1 is an original example from IcePaHC where we show a matrix clause (IP-MAT) with a subject (NP-SBJ), a main verb (VB) in the past tense (D), indicative mood (I), i.e. VBDI, an indirect object (NP-OB2), a direct object (NP-OB1) and a prepositional phrase (PP); the sentence is glossed and translated in (1). This rich annotation scheme makes the conversion to a dependency-based scheme possible.

(1)	Húsbændur-nir	borguðu	honum	vatnsburð-inn	á	vissum	tímum.
	masters.of.house-the.NOM	paid	him.DAT	water.carrying-the.ACC	at	certain	times

‘The masters of the house paid him for carrying the water at certain times.’

```
( (IP-MAT (NP-SBJ (NS-N Húsbændur$-húsbóndi) (D-N $nir-hinn))
    (VBDI borguðu-borga)
    (NP-OB2 (PRO-D honum-hann))
    (NP-OB1 (N-A vatnsburð$-vatnsburður) (D-A $inn-hinn))
    (PP (P á-á)
        (NP (ADJ-D vissum-viss) (NS-D tímum-tími)))
    (.  .-.))  (ID 1883.VOGGUR.NAR-FIC,.37))
```

Figure 1: An example of the IcePaHC format.

The practicality of using a historical (diachronic) corpus for conversion to a descriptive dependency corpus for Icelandic hinges on the fact that Icelandic has changed much less than many other European languages over the last thousand years, syntax included (Rögnvaldsson and Helgadóttir, 2011). This fact by itself supports the use of IcePaHC in the conversion. However, even though Icelandic has changed less than some other languages, it has in fact gone through various syntactic changes. These include changes from OV to VO word order, the emergence of the first position expletive *það* ‘it, there’ and decreased use of certain types of empty arguments (see Rögnvaldsson, 2005 for an overview).

3 The PPCHE and UD Annotation Schemes

The PPCHE and UD annotation schemes reflect two different ways of parsing a sentence. While the PPCHE scheme denotes phrases by using brackets, the UD scheme connects each token to a single head (Nivre et al., 2016). Every sentence in UD has a root and each dependent relationship a label, but neither are marked in the PPCHE scheme. An obvious difference between the two schemes is how information on a sentence's annotation is displayed. Figure 2 displays the dependency-based counterpart

[5] A description of the tagset can be found at `https://linguist.is/icelandic_treebank/Tagset`

to the IcePaHC sentence in Figure 1. Comparing the two, we see that the PPCHE format focuses on phrases while the UD format, which is delivered in CoNLL-U format, lists all tokens in a sentence and gives information on them. The CoNLL-U format consists of ten fields, or columns, as outlined below.

```
ID  FORM          LEMMA        UPOS   XPOS   FEATS                   HEAD  DEPREL  DEPS  MISC
1   Húsbændurnir  húsbóndi     NOUN   NS-N   Case=Nom|Number=Plu...  2     nsubj   _     IFDtag=nkfng
2   borguðu       borga        VERB   VBDI   Number=Plur|Mood=In...  0     root    _     IFDtag=sfg3fþ
3   honum         hann         PRON   PRO-D  Case=Dat|Number=Sin...  2     iobj    _     IFDtag=fpkeþ
4   vatnsburðinn  vatnsburður  NOUN   N-A    Case=Acc|Number=Sin...  2     obj     _     IFDtag=nkeog
5   á             á            ADP    P      AdpType=Prep|Degree...  7     case    _     IFDtag=aþ
6   vissum        viss         ADJ    ADJ-D  Case=Dat|Number=Plu...  7     amod    _     IFDtag=lkfþsf
7   tímum         tími         NOUN   NS-D   Case=Dat|Number=Plu...  2     obl     _     IFDtag=nkfþ|
                                                                                         SpaceAfter=No
8   .             .            PUNCT  .      _                       2     punct   _     .
```

Figure 2: An example of the output CoNLL-U format, taken from the Icelandic UD corpus.

- The 1st, 2nd, and 3rd columns contain the word index (ID), the word form (FORM), and the lemma (LEMMA), respectively.
- The 4th column contains the Universal PoS tags (UPOS). This is the UD format tag corresponding to the tag in column 5 and it is discussed further in Section 4.2.
- The 5th column contains the language-specific PoS tags as used in IcePaHC (XPOS).
- The 6th column contains morphological features (FEATS), which give detailed information on a word beyond its UPOS and XPOS tags, for example case, number, tense, voice and person, all depending on word classes. These are discussed further in Section 4.2.
- The 7th column (HEAD) specifies which word the current word is dependent on, by using the head's word index. Each sentence has one root, specified with a '0', on which all other words in the sentence are dependent, whether the relation is direct or through other words.
- The 8th column contains the code for the universal dependencies relation (DEPREL), which states of what kind the relation between a word and its head is, discussed in Section 4.3.
- The 9th column is left blank but it can include information on enhanced dependencies (DEPS).
- Finally, there is a miscellaneous column (MISC). In the case of the Icelandic UD corpus, this column is used for displaying the word's revised Icelandic Frequency Dictionary (IFD) (Pind et al., 1991) morphosyntactic PoS tag, given by the ABLTagger (Steingrímsson et al., 2019), as discussed in Section 4.2.

4 The Conversion

Our method of converting IcePaHC to a UD corpus mainly consists of three steps. The first step involves text cleanup, i.e. removing information not critical to UD. Next is the actual conversion, extracting information on the words and phrase structure and delivering it according to the CoNLL-U format. Lastly, various post-processing is done on the CoNLL-U output files, in order to meet UD format standards. The converter and its resulting treebank have not been evaluated formally but numerous sentences have been checked manually while developing the converter. Systematic evaluation will be carried out in future work.

4.1 Text Cleanup

Since the phrase structure of the sentences in IcePaHC is depicted using brackets, the NLTK CategorizedBracketParseCorpusReader[6] is used. In order for it to operate, some cleanup needs to be carried out on the corpus files. This mostly consists of removing additional information included in the files, e.g. sentence ID tags and the annotators' notes, along with empty lines and extra brackets. In the annotation

[6] https://www.nltk.org/_modules/nltk/corpus/reader/bracket_parse.html

scheme, various words that are normally written as one are split into separate tokens, and as the UD format does not follow suit, this process has to be reversed. This includes suffixed articles, e.g. *-in* in *mær-in* 'maiden-the'. These are split from the noun in IcePaHC as separate tokens but appear as a part of the noun in the UD scheme. In addition to these changes, a specific script fixes various minor annotation errors in the corpora themselves, discovered while the converter was being developed.

The second part of this cleanup stage is carried out after the conversion itself, in which the output CoNLL-U files are modified. For example, we transparently portray the cliticization of pronouns to their corresponding verbs in the CoNLL-U output. The annotation scheme splits these pronouns or pronoun clitics from their corresponding verbs such that, e.g., *heyrðu* (imperative 'hear (you)') is split into the imperative form *heyr* 'hear' and the second person singular pronoun clitic *-ðu* 'you'. Keeping consistent with UD annotation guidelines (e.g., for Spanish and German), we adopt the annotation scheme shown in Figure 3, where the word components are shown both combined (ID 1–2) and split (IDs 1 and 2), so that the surface form is apparent but all features and dependency relations of the components are also clear. Included in this last phase of the conversion pipeline is joining together sentences so that a sentence in the final output is defined by a full stop rather than by a matrix clause, to adhere to UD convention.

```
ID   FORM    LEMMA  UPOS
1-2  Heyrðu  _      _
1    Heyr    heyra  VERB
2    þú      þú     PRON
```

Figure 3: An example of a verb-clitic relation marked by index range (further CoNLL-U fields omitted).

4.2 Part-of-Speech Tags and Morphological Features

Information on PoS tags and morphological features is extracted in two ways. To obtain the UD tag, a word's original tag from IcePaHC is used along with handwritten rules, which map each original tag to a corresponding UD tag. The tagset used in the converted corpus consists of 17 tags, displayed in Table 1.

Tags	Description	Tags	Description	Tags	Description
ADJ	adjective	INTJ	interjection	PROPN	proper noun
ADP	adposition	NOUN	noun	SCONJ	subordinating conjunction
ADV	adverb	NUM	numeral	SYM	symbol
AUX	auxiliary verb	PART	particle	VERB	verb
CCONJ	coordinating conjunction	PUNCT	punctuation	X	other
DET	determiner	PRON	pronoun		

Table 1: The UD tags used in the converted corpus.

The mapping from an IcePaHC tag to a UD tag is not always unequivocal since the tagging scheme used in the corpus groups some word classes together, which are kept separate in the UD tagging scheme, and vice versa. An example of this variation is the treatment of quantifiers. They are not included in the PoS categories used in UD but they are included in the IcePaHC tagset, and so an appropriate tag has to be chosen. Some quantifiers are classified as adjectives in UD and some as pronouns, but this distinction is not marked in the original corpus and therefore one tag has to be chosen for all. Since the majority of quantifiers are categorized as adjectives, a decision was made to map quantifiers to adjectives. Another problem of a similar kind is that verbs in IcePaHC can have the same tag whether they are a main or an auxiliary verb, but in UD the tags are different. For example, the Icelandic verb *hafa* 'have' can both act as a main verb, as in (2), and auxiliary verb, as in (3), but it is in both cases tagged as HV in IcePaHC. Therefore, a correct tag has to be chosen based on its context at any given time.

(2) Hún **hafði** lítinn tíma.
she had little time

(3) Hún **hafði** notað lítinn tíma.
she had used little time

Although it is thorough, the IcePaHC PoS tagset does not include enough information to display the complete morphological features for all words in the converted corpus, as various grammatical features are omitted from the annotation. For example, grammatical gender (masculine, feminine and neuter) is not present in the tagset, which affects, e.g., nouns and adjectives. This means that a full, automatic conversion from IcePaHC tags to the UD feature scheme, which correctly describes the grammatical features of Icelandic, is not possible in a simple step.

To obtain more information on the tokens, the IcePaHC PoS tags are skipped altogether and the whole text of the treebank is extracted and automatically PoS-tagged using ABLTagger, a state-of-the-art PoS tagger for Icelandic (Steingrímsson et al., 2019). It returns IFD-format PoS tags, mentioned in Section 3, from which morphological features are extracted. The output of this step, although not hand-tagged, is considerably more detailed than the IcePaHC tags, thus providing all necessary morphological features for the output UD corpus, as opposed to the incomplete IcePaHC PoS tags.[7] The tags from the ABLTagger are shown in the miscellaneous (MISC) column in the CoNLL-U format in Figure 2.

4.3 Heads and Dependency Relations

Extracting relational information from the IcePaHC format involves selecting each phrase's head and its dependents and the type of dependency relation. For this, we build on an experimental project carried out by Örvar Kárason.[8] The method of conversion is based on similar work originally done by Magerman (1994) and improved by others (Yamada and Matsumoto, 2003; Johansson and Nugues, 2007).

Each sentence in the given corpus is read with the help of NLTK[9] and information on tags from the original files is also used. A matrix clause is read bottom-up to determine the head of all its subordinate phrases, skipping traces and empty nodes. This head selection is done with handwritten rules that match a phrase tag with its possible heads, which in turn are ordered according to priority so that a correct head is chosen if more than one are possible. The IcePaHC PoS tags are suitable for this task and as they have been manually corrected, no information on phrase type is lost during conversion between tagsets. By selecting phrasal heads bottom-up, we establish a hierarchy where every word or phrase within a phrase is dependent on the head, culminating in the head of the whole sentence, which, according to UD guidelines, is the sentence's root.

When determining the type of dependency relation between two words, information on a phrasal head is used along with the words' PoS tags. Handwritten rules are used to match a tag with its dependency relation, some of which specify the phrasal head's tag to handle ambiguous cases. Again, the detailed tagset of IcePaHC makes this process possible. The dependency relations used can be seen in Table 2. All dependency relations used were crucial for the conversion to be precise and, in most cases, phrasal heads in IcePaHC correspond to a dependency relation. An exception to this are the relations 'acl', 'advcl', 'ccomp' and 'xcomp'. IcePaHC does not make a distinction between 'acl' and 'advcl' on the one hand and 'ccomp' and 'xcomp' on the other, but UD does. These dependency relations are therefore handled specifically and the appropriate relation chosen based on the sentence it appears in.

Not only are the dependency relations important, but their direction is too. Some relations can only go from left to right and others from right to left, and this had to be considered when creating the converter since no equivalent rule is present in IcePaHC. In some cases, the original annotation is changed to adhere to UD rules, for example when a phrase includes two auxiliary verbs, one of them being the phrasal head. The other auxiliary verb cannot be dependent on another auxiliary verb and the relation must therefore be changed, so that both auxiliary verbs are dependent on a main verb.

[7] This approach has various issues in itself, both in execution, as it relies on external software, and in output accuracy, as the PoS tagger used is considered state-of-the-art for modern Icelandic, but presumably performs worse as the input texts deviate more from the standard language and orthography. This will eventually be hand-checked and evaluated.

[8] `https://github.com/OKarason/venzl`

[9] `https://www.nltk.org/_modules/nltk/corpus/reader/bracket_parse.html`

Dependency relations	Description	Dependency relations	Description
acl	adjectival clause	dislocated	dislocated elements
acl:relcl	relative clause modifier	expl	expletive
advcl	adverbial clause modifier	fixed	fixed multiword expression
advmod	adverbial modifier	flat:foreign	foreign words
amod	adjectival modifier	flat:name	multiword expression with name
appos	appositional modifier	iobj	indirect object
aux	auxiliary	mark	marker
case	case marking	nmod	nominal modifier
cc	coordinating conjunction	nmod:poss	possessive nominal modifier
ccomp	clausal complement	nsubj	nominal subject
compound	compound	nummod	numeric modifier
compound:prt	phrasal verb particle	obj	object
conj	conjunct	obl	oblique nominal
cop	copula	obl:arg	oblique argument
csubj	clausal subject	parataxis	parataxis
dep	unspecified dependency	punct	punctuation
det	determiner	vocative	vocative
discourse	discourse element	xcomp	open clausal complement

Table 2: Dependency relations used in the converted corpus.

5 Extension: Converting a Faroese Treebank

The goal of this project is not to produce a one-shot converter for a specific treebank, but a reusable tool, applicable to treebanks in a similar format. Icelandic and Faroese share a number of linguistic properties that make constituency grammar annotation using the same scheme possible. To demonstrate the reusability of the tool, it was adapted to the Faroese Parsed Historical Corpus (FarPaHC)[10] which uses the same annotation scheme as IcePaHC. It consists of three New Testament translations, i.e. one from the 19th century and two from the 20th century, 53,000 words in total. The 19th century text has not been included in the converted UD treebank yet. Its orthography deviates in many ways from modern spelling; even though the text was modernized for FarPaHC, we decided that before we would add it to the converted treebank, more work on fixing inconsistencies would be needed. A few minor additions were needed for the conversion to be possible, which shows the utility of the conversion pipeline.

The annotation process for FarPaHC is the same as that for IcePaHC; both of them have been manually corrected according to the PPCHE annotation scheme, with some changes to adapt it to Icelandic and Faroese grammar. The tagset used in FarPaHC is for the most part the same as in IcePaHC, which is possible because of the similarities in the languages' grammars. The main difference in the annotation scheme between the two corpora is that lemmas are not shown in FarPaHC. An example of a parsed sentence in FarPaHC is shown in Figure 4, which shows a matrix clause (IP-MAT) with a direct object (NP-OB1), an auxiliary verb (HVPI), a subject (NP-SBJ), a main verb in the past participle (VBN) and two prepositional phrases (PP); the sentence is glossed and translated in (4).

[10] `https://github.com/einarfs/farpahc`

(4) Hetta havi eg talað til tykkara í líknilsum.
this have I said to you.PL in figures
'I have spoken to you about this figuratively.'

```
( (IP-MAT-SPE (CODE VS:XVI_25J)
    (NP-OB1 (D-A Hetta))
    (HVPI havi)
    (NP-SBJ (PRO-N eg))
    (VBN talað)
    (PP (P til)
        (NP (PRO-G tykkara)))
    (PP (P í)
        (NP (NS-D líknilsum)))
    (. .-.)) (ID 1936.NTJOHN.REL-BIB,.1310))
```

Figure 4: An example of the FarPaHC format.

Since the two treebanks share an annotation scheme and a tagset, FarPaHC was converted using the methods described in Section 4 with only three modifications. First, the script used to fix various annotation errors in IcePaHC was not used as it did not apply to FarPaHC. No such errors were encountered and such a script was therefore unnecessary. Second, since FarPaHC does not include information on lemmas, neither does the resulting UD corpus. Third, morphological features for FarPaHC were extracted in a different manner. The conversion from IcePaHC relies on an Icelandic high-accuracy PoS tagger for additional information on tokens but no such tagger exists for Faroese. Thus, no third-party software was used for feature extraction and all information was retrieved from the PoS tags themselves. Every component in a tag was mapped to a corresponding morphological feature using handwritten rules, which does not provide all possible features for a given token but does provide more details than solely the tag. These rules can then be built upon if the conversion pipeline is extended for another language with a similar tagset.

The conversion results in a new Faroese UD corpus of 40,000 words to be included in the next UD release. Figure 5 shows the sentence from Figure 4 after having been converted to CoNLL-U. As mentioned, lemmas are not shown, with the exception of the punctuation mark, fewer morphological features are shown and no additional tags are shown in the MISC column. In other respects, all grammatical information represented in the converted Icelandic UD corpus is also represented in the Faroese one.

```
ID  FORM       LEMMA  UPOS   XPOS   FEATS                             HEAD  DEPREL  DEPS  MISC
1   Hetta      _      DET    D-A    Case=Acc                          4     obj     _     _
2   havi       _      AUX    HVPI   Mood=Ind|Tense=Pres               4     aux     _     _
3   eg         _      PRON   PRO-N  Case=Nom                          4     nsubj   _     _
4   talað      _      VERB   VBN    Tense=Past|VerbForm=Part          0     root    _     _
5   til        _      ADP    P      _                                 6     case    _     _
6   tykkara    _      PRON   PRO-G  Case=Gen                          4     obl     _     _
7   í          _      ADP    P      _                                 8     case    _     _
8   líknilsum  _      NOUN   NS-D   Case=Dat|Definite=Ind|Num...      4     obl     _     SpaceAfter=No
9   .          .      PUNCT  .      _                                 4     punct   _     _
```

Figure 5: A sentence converted to UD from FarPaHC.

6 Future Development and Use

IcePaHC currently consists of about one million words but 75,000 additional words were annotated in this project by Kristján Rúnarsson in accordance with the IcePaHC scheme (Rúnarsson and Sigurðsson, 2020). This is done to increase the weight of modern Icelandic in the corpus. These additions, which contain genres not found previously in IcePaHC – parliamentary speeches and sports news texts taken from the Icelandic Gigaword Corpus (Steingrímsson et al., 2018) – will be converted to the UD scheme using the conversion tool described in this paper, making for a close to 1.1 million word corpus. Not only does this enlarge the corpus, but also increases the ratio of modern texts in it.

A possible improvement to the Faroese UD treebank is adding information on lemmas. Since FarPaHC does not include lemmas, a Faroese lemmatizer would have to be used. A few lemmatizers are available for this task (Yildiz and Tantuğ, 2019; Kanerva et al., 2018; Rosa and Mareček, 2018), which would have to be evaluated further in regards to their feasibility.

Compatibility between the current project and the IcePaHC and FarPaHC corpora means that various possibilities exist for future interlinking of projects that make use of the two resources. Firstly, by using the conversion tool for both an Icelandic and a Faroese treebank, we have shown that cross-lingual use is possible. Secondly, with the possibility of user customization, e.g. specifying language-specific rules and parameters, the converter may be applied – with necessary modifications – to various other constituency treebanks to produce new UD corpora for a wide range of languages. These include corpora of Historical Greek (Beck, 2011), Historical Portuguese (Galves, 2018) and Middle Low German (Booth et al., 2020), in addition to PPCHE. Finally, the conversion tool can also be used to convert output from automatic parsers which use the IcePaHC annotation scheme, e.g. IceNeuralParsingPipeline (Arnardóttir and Ingason, 2020), to UD format. Since the original corpus would be automatically created, the UD format output would have to be manually corrected before adding it to the Icelandic UD corpus.

7 Conclusion

In this paper, we have described a rule-based conversion pipeline applicable to the Icelandic Parsed Historical Corpus and the Faroese Parsed Historical Corpus, which results in corpora annotated according to Universal Dependencies. The pipeline is used for converting the two treebanks, delivering a 1 million word Icelandic UD corpus and a 40,000 word Faroese UD corpus. The Icelandic corpus will consequently be among the largest of its type.

We have discussed the benefits of a UD-based corpus for Icelandic and Faroese and why IcePaHC is a good candidate for building the conversion on. The process of converting sentences was described and the methodology behind it, extracting information needed for a UD corpus by using both information included in the IcePaHC annotation scheme and a state-of-the-art PoS tagger for Icelandic. We also discussed some challenges faced because of the different annotation scheme of the original treebank and UD and how these were dealt with. We demonstrated the conversion pipeline's applicability to different languages by converting a Faroese constituency treebank to a UD corpus using the pipeline. Lastly, we discussed further possibilities in using the converter and the UD corpora with available language resources along with the 75,000 word addition to IcePaHC, which will be converted using the pipeline and added to the Icelandic UD corpus.

The benefits of an Icelandic Universal Dependencies corpus are significant, as the UD annotation scheme offers a standard parsing framework across different languages. The information that is annotated in IcePaHC is a superset of what is required by the UD scheme and this makes conversion between the two formats a feasible approach to creating a UD treebank for Icelandic. Furthermore, the open-source policy of IcePaHC as well as the global UD project is important and therefore maintained in this project. The conversion tool is released under an Apache 2.0 license on GitHub[11] and the two UD corpora are included in the next release of UD under a CC BY-SA 4.0 license.

8 Acknowledgements

This project was funded by the Strategic Research and Development Programme for Language Technology, grant no. 180020-5301. The conversion of the Faroese corpus was funded by the Icelandic Student Innovation Fund, grant no. 206457-0091. Thanks are due to Örvar Kárason, whose work was used as a basis for the conversion. We also thank Kristján Rúnarsson for helpful discussions. Finally, we would like to thank two anonymous reviewers for their helpful comments.

[11] `https://github.com/thorunna/UDConverter`

References

Þórunn Arnardóttir and Anton Karl Ingason. 2020. A neural parsing pipeline for Icelandic using the Berkeley neural parser. In Costanza Navarretta and Maria Eskevich, editors, *Proceedings of CLARIN 2020*, pages 48–51.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Hannah Booth, Anne Breitbarth, Aaron Ecay, and Melissa Farasyn. 2020. A Penn-style treebank of Middle Low German. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 766–775, Marseille, France. European Language Resources Association.

Charlotte Galves. 2018. The Tycho Brahe Corpus of Historical Portuguese: Methodology and results. *Linguistic Variation*, 18(1):49–73.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *Proceedings of Sixth International Conference on Natural Language Processing*, GoTAL 2008, pages 205–216, Gothenburg, Sweden.

Anton Karl Ingason, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel C. Wallenberg. 2012. Faroese Parsed Historical Corpus (FarPaHC). Version 0.1.

Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel Wallenberg. 2014. Rapid deployment of phrase structure parsing for related languages: A case study of Insular Scandinavian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 91–95, Reykjavík, Iceland. European Language Resources Association.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 105–112, Tartu, Estonia. University of Tartu, Estonia.

Hildur Jónsdóttir and Anton Karl Ingason. 2020. Creating a parallel Icelandic dependency treebank from raw text to Universal Dependencies. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 2924–2931, Marseille, France. European Language Resources Association.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142, Brussels, Belgium. Association for Computational Linguistics.

Anthony S. Kroch and Ann Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second edition. Size: 1.3 million words.

Anthony S. Kroch, Beatrice Santorini, and Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition. Size: 1.8 million words.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of Interspeech – Speech and language technology for less-resourced languages*, Interspeech 2007, Antwerp, Belgium.

David Mitchell Magerman. 1994. *Natural Language Parsing As Statistical Pattern Recognition*. Ph.D. thesis, Stanford University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Paris, France. European Language Resources Association.

Jörgen Pind, Friðrik Magnússon, and Stefán Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavík.

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2011. Morphological tagging of Old Icelandic texts and its use in studying syntactic variation and change. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 63–76, Berlin. Springer.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2011. Creating a dual-purpose treebank. *JLCL*, 26(2):139–150.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1977–1984, Istanbul, Turkey. European Language Resource Association.

Eiríkur Rögnvaldsson. 2005. Setningafræðilegar breytingar í íslensku. In Höskuldur Thráinsson, editor, *Setningar.* Handbók um setningafræði. Íslensk tunga III, pages 602–635. Almenna bókafélagið, Reykjavík.

Rudolf Rosa and David Mareček. 2018. CUNI x-ling: Parsing under-resourced languages in CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 187–196, Brussels, Belgium. Association for Computational Linguistics.

Kristján Rúnarsson and Einar Freyr Sigurðsson. 2020. Parsing Icelandic Alþingi transcripts: Parliamentary speeches as a genre. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 44–50, Marseille, France. European Language Resources Association.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of RANLP 2019*, Varna, Bulgaria.

Francis M. Tyers, Mariya Sheyanova, Alexandra Martynova, Pavel Stepachev, and Konstantin Vinogradovsky. 2018. Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebank.

Hiroyasu Yamada and Yuji Matsumoto. 2003. Statistical Dependency Analysis with Support Vector Machines. In *The 8th International Workshop of Parsing Technologies (IWPT2003)*, pages 195–206, Nancy, France.

Eray Yildiz and A. Cüneyd Tantuğ. 2019. Morpheus: A neural network for jointly learning contextual lemmatization and morphological tagging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 25–34, Florence, Italy, August. Association for Computational Linguistics.