*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2924–2931
Marseille, 11–16 May 2020


# Creating a Parallel Icelandic Dependency Treebank from Raw Text to Universal Dependencies

**Hildur Jónsdóttir, Anton Karl Ingason**
University of Iceland
Sæmundargötu 2, 101 Reykjavík
hildur.jonsdottir@gmail.com, antoni@hi.is

**Abstract**

Making the low-resource language, Icelandic, accessible and usable in Language Technology is a work in progress and is supported by the Icelandic government. Creating resources and suitable training data (e.g., a dependency treebank) is a fundamental part of that work. We describe work on a parallel Icelandic dependency treebank based on Universal Dependencies (UD). This is important because it is the first parallel treebank resource for the language and since several other languages already have a resource based on the same text. Two Icelandic treebanks based on phrase-structure grammar have been built and ongoing work aims to convert them to UD. Previously, limited work has been done on dependency grammar for Icelandic. The current project aims to ameliorate this situation by creating a small dependency treebank from scratch. Creating a treebank is a laborious task so the process was implemented in an accessible manner using freely available tools and resources. The parallel data in the UD project was chosen as a source because this would furthermore give us the first parallel treebank for Icelandic. The Icelandic parallel UD corpus will be published as part of UD version 2.6.

**Keywords:** low-resource languages, parallel treebanks, Universal Dependencies, Icelandic

## 1. Introduction

In order to survive the competition with a global English in various technology-associated domains, the Icelandic language, spoken by 350,000 people, must meet the challenges brought on by developments in Language Technology. Although it is not yet considered to be in imminent danger (Rögnvaldsson et al., 2012b), a number of efforts are currently underway to address this situation. One of the core projects that the Icelandic government is supporting to achieve this is to build treebanks and especially dependency treebanks (Nikulásdóttir et al., 2017). In recent years, The Universal Dependencies project (Nivre et al., 2016) has been a leading force in parsing and cross-lingual research and becoming a part of it can make Icelandic Language Technology more viable. A treebank based on this type of an annotation scheme could become a foundation for further Icelandic parser development because treebanks are the essential training data for natural language data-driven parsers. The widespread interest that the UD project has received may also generate more interest in working on Icelandic Language Technology solutions in general once Icelandic UD resources are available.

No dependency parser has yet been developed for Icelandic. However three phrase structure parsers are available. These are IceParser, a shallow phrase-structure parser which is a part of the IceNLP toolkit (Loftsson and Rögnvaldsson, 2007), Greynir, a rule-based parser based on context-free grammar (Þorsteinsson et al., 2019), and a parsing pipeline built on the IcePaHC treebank and the Berkeley Parser (Jökulsdóttir et al., 2019). A parser for Icelandic could for example support development of an Icelandic grammar checker and be useful in applications like question answering, machine translation, information extraction and speech generation/understanding (Nikulásdóttir et al., 2017). Since previous work on dependency grammar for Icelandic is sparse we decided to start with studying the UD annotation scheme by working on a small corpus from scratch. As the core of the UD project is about consistency and parallelism, we focused on adjusting the annotation scheme to related languages[1] without sacrificing any elements.

At the same time as the present project on a parallel treebank took place, another team carried out work on a conversion tool from the IcePaHC treebank (Rögnvaldsson et al., 2012a) to UD. IcePaHC is an Icelandic treebank based on the annotation scheme for the Penn Parsed Corpora for Historical English. We collaborated on finding the best solutions for a shared Icelandic annotation documentation. It has been shown that converted treebanks are missing rare constructions that original treebanks feature (Peng and Zeldes, 2018) so this work was helpful in developing the Icelandic annotation scheme. The parallel corpora in UD (PUD) are based on 1,000 sentences from newspaper texts and Wikipedia which is a genre that is not part of the IcePaHC corpus. We consider this to be a valuable choice of data because it delivers a parallel corpus with accurate 1-1 sentence alignment for 19 other languages (Nivre et al., 2019). This first Icelandic parallel treebank will be freely available on Github[2] and the process for creating it is described in this paper. We begin with the raw data which needed a good translation and then our journey through automatic tagging, lemmatizing, conversion to CoNLL-U format, preprocessing the syntactic annotation with delexicalized methods, manual correction and evaluation.

## 2. Related work

With the growing demand on resources for natural language processing (NLP), the first Icelandic treebank came to light in 2011 (Rögnvaldsson et al., 2012a). The parsing scheme

[1] Henceforth, the North Germanic languages; Danish, Faroese, Norwegian and Swedish

[2] `https://github.com/UniversalDependencies/UD_Icelandic-PUD/`

was originally designed for the Penn Parsed Corpora for Historical English and it uses phrase structure annotation in a labelled bracketing format. At the same time, dependency treebanks were being built for related languages. However, because the Penn scheme is quite detailed, it contains the information required to convert it to dependency grammar but not vice versa. Holding 1 million tokens and spanning almost 10 centuries, the purpose of IcePaHC is twofold, to be suitable for both language technology and syntactic research. Another treebank based on wide-coverage context free grammar is being developed (Þorsteinsson et al., 2019). The plan is to convert it to dependency annotated corpora for training deep neural network-based parsers. Other Icelandic corpora suited for NLP have been growing steadily in the last decades[3]. To be mentioned here is the Icelandic Gigaword corpus (IGC) (Steingrímsson et al., 2018), a corpus of about 1,300 million words, tagged with the IFD tagset described in 3.3. It mainly holds web media and printed papers. Another notable corpus recently published is the first English-Icelandic parallel corpus for the purposes of language technology development and research, ParIce (Barkarson and Steingrímsson, 2019). It consists of 38.8 million words in 3.5 million segmented pairs automatically aligned. The main purpose of this corpus is for training machine translation systems but could also be used for, e.g., creating dictionaries and ontologies, multilingual and cross-lingual document classification.

It is important to review the work done for related languages in UD because the project focuses on cross-lingual studies. There are pros and cons in being the last North Germanic language to participate in the UD project. The annotation scheme has been improved since the first version and multiple tools have been developed to ease the tasks. The apparent disadvantage is that the Icelandic language has not been a part of the UD studies, so far. The first public dependency treebank for Norwegian Bokmål and Nynorsk, The Norwegian Dependency Treebank (NDT), was published in 2014 (Solberg et al., 2014) and later converted to UD (Øvrelid and Hohle, 2016). The NDT annotations were made with consideration to similar treebanks, the Swedish treebank Talbanken and the treebank of old Indo-European languages PROIEL. The corpus is divided into Bokmål (310K tokens) and Nynorsk (301K tokens) and contains mostly newspaper texts. A UD treebank of spoken dialects is also available in Norwegian, LIA (Øvrelid et al., 2018), which was annotated with morphological and dependency-style syntactic analysis according to the LIA project and later converted to UD. The purpose of the corpus, which has 55K tokens, is to increase research on spoken Norwegian with parser development in mind.

The Danish UD treebank (Johannsen et al., 2015) is a conversion of the Danish Dependency Treebank (DDT). The DDT derives from a morphosyntactically tagged corpus created for a EU project called Parole. The texts are of various genre, mainly newspapers and the grammar of DDT is based on discontinuous grammar.

For Swedish there are three UD treebanks available. Talbanken has been a part of UD since version 1, it consists of about 95,000 tokens converted from the Swedish Talbanken (Nivre and Megyesi, 2007). It has various text genres including textbooks, information brochures and newspaper articles. Another Swedish UD treebank is LinES (Ahrenberg, 2015) which was originally designed as a parallel treebank based on dependency grammar and later converted to UD. The English source is also available on UD. The texts are of literary genre, online manuals and Europarl data and count total of about 90,000 tokens. The third Swedish treebank, Swedish-PUD, was created for the CoNLL 2017 Shared Task (Nivre et al., 2017). It is available as a test file in UD like all the Parallel Universal Dependencies treebanks.

For Faroese, which is the closest relative of Icelandic and spoken by only 72,000 people, there is a UD corpus with 10K tokens including texts from Faroese Wikipedia (Tyers et al., 2018a).

As can be seen from the above cases the creation and nature of UD treebanks varies between the related languages but most of them are a converted version of dependency based treebanks.

## 3. Data and Tools

### 3.1. Source Data

Along the conversion of IcePaHC to UD we decided to create a small corpus from scratch to reveal all elements needed and to ensure consistency and parallelism for the Icelandic annotation scheme. The source data chosen was an Icelandic version of Parallel Universal Dependencies. Parallel treebanks can be used for translation studies, as training or evaluation corpora for word or sentence alignment, input for example-based machine translation (EBMT) and as training data for transfer rules (Volk et al., 2018). Since this corpus is small it is better suited for testing and evaluation than training purposes.

The parallel corpora in UD were specially prepared for the CoNLL 2017 Shared Task (Nivre et al., 2017) and are now available in English, Swedish, French, Japanese, Polish, Turkish, Thai, Spanish, Russian, Portuguese, Korean, Italian, Indonesian, Hindi, German, Finnish, Czech, Chinese and Arabic. The shared task was about syntactic dependency parsers that work for typologically different languages by exploiting a common syntactic annotation standard. The texts are mainly from news and Wikipedia and include 1,000 sentences which map 1-1 to other PUD treebank sentences. The first 750 sentences are originally English but the remaining 250 sentences are originally German, French, Italian or Spanish and were translated to English which is the source language.

Unlike other PUD treebanks, the Icelandic PUD was not created as part of the CoNLL 2017 Shared Task. The first step was to translate the data from English to Icelandic and therefore a professional translator, Ölvir Gíslason, was recruited to translate all 1,000 sentences. He was only given the guidelines to let the sentences match accurately 1-1. The translation has not been altered in any way and gave exactly 1,000 sentences and 18,812 tokens.

[3] http://www.malfong.is/

### 3.2. Adjusting Icelandic to the UD Annotation Scheme

The UD project requires each language to share their annotation specification with other treebanks of same language to increase consistency and parallelism. The focus when adjusting Icelandic to the UD annotation scheme was on alignment with related languages without loosing any elements. The tagset designed for the IcePaHC corpus differs in some ways from the IFD tagset described in 3.3. and applied to most Icelandic corpora. In general the IcePaHC treebank holds very detailed syntactic information. However, its tagset includes less features which were added to the conversion with additional tagging. Foreign names, brands, symbols and copula sentences were more noticable in the Icelandic PUD whereas first or second person sentences and discourse elements were more frequent in IcePaHC. The Icelandic annotation utilizes all the Universal part-of-speech tags (UPOS), listed in table 1. The lexical and inflectional features chosen for the Icelandic annotation are listed in table 2, the strikethrough features were not included. All the main features are parallel with related languages, but Norwegian includes animacy and both Swedish and Danish include the foreign feature. This difference is inevitable and should be minor for most research and processing. The main difference here will be on the feature values as the inflectional morphology of Icelandic is richer than that of the other North Germanic languages. There are small variations on the dependency relations, mostly subtype relations, between the related languages and Icelandic. The obl:arg relation introduced in version 2 of UD (Zeman, 2017) which distinguishes oblique arguments from adjuncts was added to the Icelandic annotation. The orphan, dislocated, acl:cleft, aux:pass, nsubj:pass, csubj:pass and obl:agent which are in the Swedish and Norwegian relations set are not a part of the Icelandic set in this first version but might be added later. The Icelandic relation set is listed in table 3 with strikethrough relations for those not included. Enhanced dependencies are not a part of this first version but might be added later, e.g. the case information and the ellipsis which are a part of the IcePaHC annotation.

| Open class words | Closed class words | Other |
|---|---|---|
| ADJ | ADP | PUNCT |
| ADV | AUX | SYM |
| INTJ | CCONJ | X |
| NOUN | DET | |
| PROPN | NUM | |
| VERB | PART | |
| | PRON | |
| | SCONJ | |

Table 1: List of UPOS tags

| Lexical Features | Inflectional Features | |
|---|---|---|
| PronType | Gender | VerbForm |
| NumType | ~~Animacy~~ | Mood |
| Poss | ~~NounClass~~ | Tense |
| ~~Reflex~~ | Number | Aspect |
| ~~Foreign~~ | Case | Voice |
| Abbr | Definite | ~~Evident~~ |
| ~~Typo~~ | Degree | ~~Polarity~~ |
| | | Person |
| | | ~~Polite~~ |
| | | ~~Clusivity~~ |

Table 2: List of Lexical and Inflectional Features

| | Nominals | Clauses | Modifier Words | Function Words |
|---|---|---|---|---|
| **Core Arguments** | nsubj<br>obj<br>iobj | csubj<br>ccomp<br>xcomp | | |
| **Non-Core Dependents** | obl<br>obl:arg<br>vocative<br>expl<br>~~dislocated~~ | advcl | advmod<br>discourse | aux<br>cop<br>mark |
| **Nominal Dependents** | nmod<br>nmod:poss<br>appos<br>nummod | acl | amod | det<br>~~elf~~<br>case |
| **Coordination** | **MWE** | **Loose** | **Special** | **Other** |
| conj<br>cc | fixed<br>flat<br>flat:name<br>flat:foreign<br>compound<br>compound:prt | ~~list~~<br>parataxis | ~~orphan~~<br>~~goeswith~~<br>~~reparandum~~ | punct<br>root<br>dep |

Table 3: Dependency Relations

The UD annotation scheme offers wide range of elements and it would be very interesting to add many of them. To mention here is further distinction of expletives (Bouma et al., 2018) which have received attention in Icelandic syntax studies (Árnadóttir et al., 2011). The expletive subtypes were not added this time since neither the default tagset nor the treebanks to be converted include the distinctions required.

### 3.3. Tagging

For the properties of the part-of-speech tags and features, the Icelandic translation had to be tagged. The state–of–the–art ABLTagger was used which is based on BiLSTM models, a morphological lexicon and lexical category identification (Steingrímsson et al., 2019). It is trained on texts tagged with the IFD tagset which consists of 565 tags (Loftsson et al., 2009) that has been the tagset featuring the majority of Icelandic corpora built in the last years. The Icelandic language is highly inflectional and this tagset is a combination of word classes and morphosyntactic features which makes it so large. In the CoNLL–U format used in UD, this is entirely separated, that is, Universal part–of–speech tags (UPOS) and morphological features (FEATS). The ABLTagger also tokenizes the text utilizing a tokenizer from Miðeind (Þorsteinsson et al., 2019) which greedily recognizes certain multi–token spans like dates and adverbial multi–word idioms. The training model provided is based on various texts, mainly newspaper and literature and the given accuracy is 94.17%.

### 3.4. Lemmatizing

For lemmatizing the high accuracy lemmatizer Nefnir (Ingólfsdóttir et al., 2019) was run. This lemmatizer uses tagged input and suffix substitution rules from the Database of Modern Icelandic Inflection (Bjarnadóttir et al., 2019). It

reaches accuracy of 99.55% with verified tagged input, and for text tagged with a PoS tagger, the accuracy obtained is 96.88%. The lemmas were an important input in the conversion phase, in particular for recognizing auxiliaries from other verbs and coordinating from subordinating conjunctions.

## 4. Conversion to CoNLL-U format

### 4.1. Processing UPOS and Features

The conversion from IFD tags and lemmas to UPOS and features was direct with few exceptions. Auxiliary verbs are all tagged as verbs so only the lemmas *vera* 'be', *munu* 'will' and *skulu* 'shall' were automatically converted to AUX. Other auxiliaries exist but they can also behave as non-auxiliaries so they were manually corrected. The second thing is that all indefinite, demonstrative, interrogative and possessive pronouns are tagged with pronoun tag in the original tagset. This is not as specified by the UD guidelines where these forms are tagged as determiner (DET) when they modify a noun. This was corrected in the manual process on the UPOS level but information on the pronoun is kept with the PronType feature. To maintain the parallelism to related languages, the tags of the participles, both past and present, were converted to UPOS adjective tag but the features hold information on the verb participle and therefore no information is lost. The CoNLL-U format holds 10 fields and an empty line between sentences. An example from the Icelandic PUD is given in figure 1. The ID is the index of the token in the sentence, FORM is the word form, LEMMA is the lemma , UPOS is the Universal part-of-speech tag derived from XPOS, XPOS is the language specific part-of-speech tag, here the IFD tag provided by ABLTagger, FEATS holds the morphological features, here the extracted features from XPOS and LEMMA, HEAD is the syntactic information, i.e. head of the current word, DEPREL gives the dependency relation of the HEAD, DEPS is for enhanced dependency graph and the last field, MISC is provided for any other annotation.

### 4.2. Preprocessing Syntactic Relations

Since no Icelandic dependency parser is available we decided to train a delexicalized parser to preprocess the corpus. Delexicalized parsing, which is one type of cross-lingual model transfer, was first introduced by Zeman and Resnik (Zeman and Resnik, 2008) and is nowadays considered a standard technique in cross-lingual parsing. Delexicalized models using only UPOS tags were trained with UD treebanks of related languages, Swedish, Norwegian, Danish and Faroese (Nivre et al., 2019) and tested on the first 200 sentences in the corpus which had been annotated manually from scratch with syntactic and dependency relations (HEAD and DEPREL in CoNLL-U). The parser selected for the task is UDPipe (Straka and Straková, 2017) which was on the top list of parsers in the CoNLL 2018 Shared Task on parsability[4]. This parser does not require any training or configuration for a new language and has good usability and documentation.

[4] https://universaldependencies.org/conll18/results.html

| Model | Tokens | UAS | LAS |
|---|---|---|---|
| Norwegian Nynorsk | 301,353 | 60.03% | 51.27% |
| Swedish PUD | 19,085 | 58.41% | 49.52% |
| Swedish Lines | 90,960 | 57.77% | 50.30% |
| Norwegian Bokmaal | 310,221 | 57.54% | 50.03% |
| Danish DDT | 100,733 | 56.89% | 47.32% |
| Swedish Talbanken | 96,858 | 56.71% | 48.71% |
| Faroese | 10,002 | 46.99% | 39.01% |

Table 4: Evaluation of Delexicalized Models

Interestingly, the Swedish PUD model gave the second best results (see table 4) with 58.41% accuracy in unlabeled attachment score (UAS, percentage of words with correct HEAD) and 49.52% on labeled attachment score (LAS, percentage of words with both the correct HEAD and DEPREL) consisting of only 19K tokens which can be explained by the nature of the texts being parallel. The Faroese model, which is the closest relative to Icelandic gave the lowest score as it has only 10,002 tokens. Even though the Norwegian model gave the best score we decided to train our model with the Swedish PUD data because of the small size which would give the additional corrected Icelandic data more weight in the model. The process was divided into 5 phases, increasing the Swedish PUD delexicalized model each time with 200 corrected Icelandic sentences. The first delexicalized model which consisted of the whole Swedish PUD corpus and the first 200 manually annotated sentences gave UAS score of 70.77% and LAS of 64.05% for the next test set (sentences 200-400). The last training model which held the whole Swedish PUD corpus and 800 Icelandic sentence reached 78.82% UAS and 73.78% LAS. Figure 2 is an example of a sentence perfectly parsed by the last training model.

### 4.3. Manual Correction

There are many benefits of working on an open source cross-lingual project like UD. One of them is all the available tools that are developed and are suitable for all languages. The manual correction was done with UD Annotatrix (Tyers et al., 2018b) that provides good graphical user interface for viewing and editing the annotation. The focus in the correction phase was on the syntactic and dependency relations and on the part-of-speech tags. After the manual correction the UD validation was run for automatic verification. The whole process from translation to finishing the manual correction spanned 8 weeks.

## 5. Evaluation

The quality of annotated corpora is always reflected in the final outcome of the machine learning algorithms. A standard way to evaluate the quality of corpora is using a Golden Standard Corpus (GSC)(Wissler et al., 2014). However, alternatives have to be utilized when no GSC is available. Another approach is testing the parsability so we measured the quality of the Icelandic PUD with a 10–fold cross validation. The UDPipe parser was chosen to evaluate the Icelandic PUD, the same one as used for preprocessing. The transition system "swap" was used which is a fully non–projective system and extends the projective system by

| ID | FORM | LEMMA | UPOS | XPOS | FEATS | HEAD | DEPREL | DEPS | MISC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Rúmlega | rúmlega | ADV | aa | _ | 2 | advmod | _ | _ |
| 2 | 5,7 | 5,7 | NUM | ta | NumType=Card | 3 | nummod | _ | _ |
| 3 | milljónir | milljón | NOUN | nvfn | Case=Nom\|Definite=Ind\|Gender=Fem\|Number=Plur | 7 | nsubj | _ | _ |
| 4 | Flóridabúa | Flóridabúi | PROPN | nkfe-s | Case=Gen\|Gender=Masc\|Number=Plur | 3 | nmod:poss | _ | _ |
| 5 | hafa | hafa | AUX | sfg3fn | Mood=Ind\|Number=Plur\|Person=3\|Tense=Pres\|VerbForm=Fin\|Voice=Act | 7 | aux | _ | _ |
| 6 | þegar | þegar | ADV | aa | _ | 7 | advmod | _ | _ |
| 7 | greitt | greiða | VERB | ssg | VerbForm=Sup\|Voice=Act | 0 | root | _ | _ |
| 8 | atkvæði | atkvæði | NOUN | nhfo | Case=Acc\|Definite=Ind\|Gender=Neut\|Number=Plur | 7 | obj | _ | _ |
| 9 | í | í | ADP | aþ | _ | 12 | case | _ | _ |
| 10 | tveggja | tveir | NUM | tfvfe | Case=Gen\|Gender=Fem\|Number=Plur | 11 | nummod | _ | _ |
| 11 | vikna | vika | NOUN | nvfe | Case=Gen\|Definite=Ind\|Gender=Fem\|Number=Plur | 12 | nmod | _ | _ |
| 12 | utankjörfundarkosningu | utankjörfundarkosning | NOUN | nveþ | Case=Dat\|Definite=Ind\|Gender=Fem\|Number=Sing | 7 | obl | _ | _ |
| 13 | . | . | PUNCT | . | _ | 7 | punct | _ | _ |

Figure 1: Icelandic Dependency Annotation in CoNLL-U format

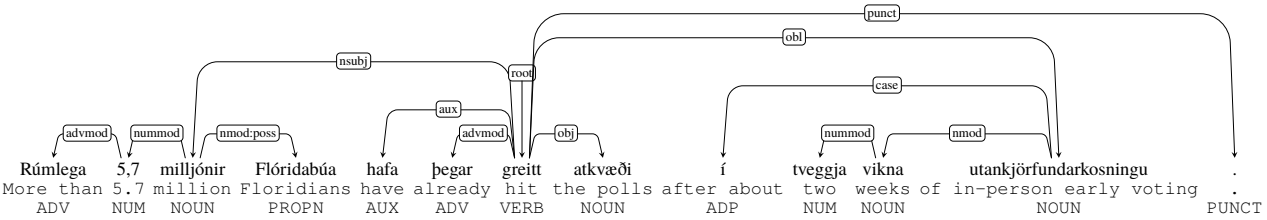


Figure 2: Icelandic Dependency Relations in Graphic Format

adding the swap transition. The transition oracle "static lazy" gives consistently better results than "static eager" according to the documentation so that was used. Other configuration was by default. We also evaluated the English, Czech and Swedish PUD for comparison, see table 5. Unsurprisingly the English PUD gives the highest score, the Swedish PUD is slightly lower and for a morphologically rich language like Czech the same model gives score about 1% above the Icelandic PUD. These measures reveal the challenges in comparing languages, even with parallel data, rather standardized text genre and accurate 1–1 sentence alignment. The Czech language is not as related to Icelandic but was evaluated here because it is morphologically rich. The Czech UD annotation scheme uses 15 UPOS tags (skips INTJ and X) and the features count 5 more than for the Icelandic UD annotation scheme. The main difference lies in the sub–features where the Czech language uses aux:pass,nsubj:pass, csubj:pass and obl:agent which the Icelandic UD annotation is missing. There is plausibly room for improvement in the Icelandic PUD corpus but as a first version we consider these results acceptable.

| Treebank | UAS | LAS |
|---|---|---|
| Icelandic PUD | 79.415% | 74.447% |
| Czech PUD | 80.45% | 75.52% |
| Swedish PUD | 82.156% | 78.65% |
| English PUD | 83.22% | 80.88% |

Table 5: 10–fold cross validation results

## 6. Conclusion

We described the first parallel treebank for Icelandic based on UD, Icelandic PUD. As a first step in studying the dependency grammar with UD annotation scheme, using the parallel data was a helpful reference to increase the parallelism desired.

Even though the preprocessing gave low accuracy compared to the best dependency parsers it definitely increased the annotation speed. For low-resource languages considering participation in the UD project we believe that the source data and method described here are simple and convenient as a first step towards UD. In our case the work was important in developing the Icelandic annotation scheme along with the conversion work for IcePaHC, especially in working with the IFD tagset and extracting the morphosyntactic features and lemmas to the Icelandic features. All new corpora to be created or converted have the option of utilizing the high accuracy ABL-tagger with the IFD tagset in order to add the features.

Although small, we hope that this corpus will be of use as part of research on the Parallel Universal Dependencies, for testing purposes and also as a reference for further development of Icelandic dependency grammar and parsing.

## 7. Bibliographical References


Ahrenberg, L. (2015). Converting an English-Swedish Parallel Treebank to Universal Dependencies. In Joakim Nivre et al., editors, Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 10–19, Uppsala, Sweden. Uppsala University.

Barkarson, S. and Steingrímsson, S. (2019). Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In Mareike Hartmann et al., editors, Proceedings of the 22nd Nordic Conference on Computational Linguistics NODALIDA-2019, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. (2019). DIM: The Database of Icelandic Morphology. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, pages 146–154, Turku, Finland. Linköping University Electronic Press.

Bouma, G., Hajic, J., Haug, D., Nivre, J., Solberg, P. E., and Øvrelid, L. (2018). Expletives in Universal Dependency Treebanks. In Marie-Catherine de Marneffe, et al., editors, Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 18–26. Association for Computational Linguistics.

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. (2019). Nefnir: A high accuracy lemmatizer for Icelandic. In Mareike Hartmann et al., editors, Proceedings of the 22nd Nordic Conference on Compu-

tational Linguistics NODALIDA-2019, pages 310–315. Linköping University Electronic Press.

Johannsen, A., Alonso, H. M., and Plank, B. (2015). Universal Dependencies for Danish. In Markus Dickinson, et al., editors, Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT14), pages 157–167.

Jökulsdóttir, T. F., Ingason, A. K., and Sigurðsson, E. F. (2019). A Parsing Pipeline for Icelandic based on the IcePaHC corpus. In K. Simov et al., editors, Proceedings of CLARIN Annual Conference 2019, Leipzig, Germany.

Loftsson, H. and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In Joakim Nivre, et al., editors, Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007, pages 128–135.

Loftsson, H., Kramarczyk, I., Helgadóttir, S., and Rögnvaldsson, E. (2009). Improving the PoS tagging accuracy of Icelandic text. In Kristiina Jokinen et al., editors, Proceedings of the 17th Nordic Conference on Computational Linguistics NODALIDA-2009, pages 103–110, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Nikulásdóttir, A., Guðnason, J., and Steingrímsson, S. (2017). Máltækni fyrir íslensku 2018-2022. Verkáætlun. Technical report, Mennta- og menningarmálaráðuneytið, Reykjavík.

Nivre, J. and Megyesi, B. (2007). Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection. In Jan Hajič Koenraad De Smedt et al., editors, Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories, pages 97–102. Northern European Association for Language Technology (NEALT).

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Paris, France. European Language Resources Association (ELRA).

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bosco, C., Bouma, G., Bowman, S., Burchardt, A., Candito, M., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cho, Y., Cinková, S., Çöltekin, Ç., Connor, M., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Droganova, K., Eli, M., Elkahky, A., Erjavec, T., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hohle, P., Ion, R., Irimia, E., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lando, T., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Moskalevskyi, B., Muischnek, K., Mustafina, N., Müürisep, K., Nainwani, P., Nedoluzhko, A., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Real, L., Reddy, S., Rehm, G., Rinaldi, L., Rituma, L., Rosa, R., Rovati, D., Saleh, S., Sanguinetti, M., Saulīte, B., Sawanakunanon, Y., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shakurova, L., Shen, M., Shimada, A., Shohibussirri, M., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Stella, A., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., van Noord, G., Varga, V., Vincze, V., Washington, J. N., Yu, Z., Žabokrtský, Z., Zeman, D., and Zhu, H. (2017). Universal Dependencies 2.0 – CoNLL 2017 Shared Task Development and Test Data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Bauer, J., Bellato, S., Bengoetxea, K., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Heinecke, J., Hennig, F., Hladká,

B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ikeda, T., Ion, R., Irimia, E., Ishola, Ọ., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kaasen, A., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Köhn, A., Kopacewicz, K., Kotsyba, N., Kovalevskaitė, J., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lam, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Li, Y., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Misirpashayeva, M., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Morioka, T., Mori, S., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perrier, G., Petrova, D., Petrov, S., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Rueter, J., Sadde, S., Sagot, B., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Särg, D., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shohibussirri, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Spadine, C., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tanaka, T., Tellier, I., Thomas, G., Torga, L., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu, H. (2019). Universal Dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Øvrelid, L. and Hohle, P. (2016). Universal Dependencies for Norwegian. In Nicoletta Calzolari, et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1579–1585, Portorož, Slovenia. European Language Resources Association (ELRA).

Peng, S. and Zeldes, A. (2018). All Roads Lead to UD: Converting Stanford and Penn Parses to English Universal Dependencies with Multilayer Annotations. In Agata Savary, et al., editors, Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 167–177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012a). The Icelandic Parsed Historical Corpus (IcePaHC). In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), pages 1977–1984, Istanbul, Turkey. European Languages Resources Association (ELRA).

Rögnvaldsson, E., Jónsdóttir, K. M., Helgadóttir, S., and Steingrímsson, S. (2012b). The Icelandic Language in the Digital Age. In Georg Rehm and Hans Uszkoreit (Eds.), *White Paper Series*. META-NET.

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannessen, J. (2014). The Norwegian Dependency Treebank. In Nicoletta Calzolari, et al., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pages 789–795, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In Nicoletta Calzolari (Conference chair), et al., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 4361–4366. European Language Resources Association (ELRA).

Steingrímsson, S., Kárason, O., and Loftsson, H. (2019). Augmenting a BiLSTM tagger with a Morphological Lexicon and a Lexical Category Identification Step. In Ruslan Mitkov, et al., editors, Proceedings of Recent Advances in Natural Language Processing, pages 1162–1169, Varna, Bulgaria.

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Jan Hajič et al., editors, Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Tyers, F. M., Sheyanova, M., Martynova, A., Stepachev, P., and Vinogradovsky, K. (2018a). Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In Marie-Catherine de Marneffe, et al., editors, Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 144–150, Brussels, Belgium. Association for Computational Linguistics.

Tyers, F. M., Sheyanova, M., and Washington, J. N. (2018b). Ud Annotatrix: An annotation tool for Universal Dependencies. In Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 10–17.

Volk, M., Marek, T., and Samuelsson, Y. (2018). Annotation, exploitation and evaluation of parallel corpora: Tc3 i. In Silvia Hansen-Schirra, Stella Neumann and Oliver Čulo (Eds.), *Translation and Multilingual Natural Language Processing 3*. Language Science Press, Berlin.

Wissler, L., Almashraee, M., Monett, D., and Paschke, A. (2014). The Gold Standard in Corpus Annotation. In 5th IEEE Germany Student Conference, IEEE GSC 2014, June 26-27, 2014, Passau, Germany. IEEE.

Zeman, D. and Resnik, P. (2008). Cross-Language Parser Adaptation between Related Languages. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

Zeman, D. (2017). Core Arguments in Universal Dependencies. In Simonetta Montemagni et al., editors, Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy, pages 287–296. Linköping University Electronic Press.

Árnadóttir, H., Eythorsson, T., and Sigurðsson, E. (2011). The passive of reflexive verbs in Icelandic. *Nordlyd*, 37:39–97.

Øvrelid, L., Kåsen, A., Hagen, K., Nøklestad, A., Solberg, P. E., and Johannessen, J. B. (2018). The LIA Treebank of Spoken Norwegian Dialects. In Nicoletta Calzolari (Conference chair), et al., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 4482–4488. European Language Resources Association (ELRA).

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. (2019). A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System. In Ruslan Mitkov, et al., editors, Proceedings of Recent Advances in Natural Language Processing, pages 1397–1404, Varna, Bulgaria.