A Parallel Icelandic Dependency Treebank: Creation, Annotation and Evaluation

Hildur Jónsdóttir
kt.170377-3389

A PARALLEL ICELANDIC DEPENDENCY TREEBANK: CREATION, ANNOTATION AND EVALUATION

Advisor
Anton Karl Ingason
Assistant Professor at The University of Iceland

Faculty of Icelandic and Comparative Cultural Studies
School of Humanities
University of Iceland
Reykjavik, January 2020

A Parallel Icelandic Dependency Treebank: Creation, Annotation and Evaluation
30 ECTS thesis submitted in partial fulfillment of a M.A. degree in Language Technology

Faculty of Icelandic and Comparative Cultural Studies
School of Humanities
University of Iceland
Sæmundargata 2
102, Reykjavik
Iceland

Telephone: +354 525 4400

Bibliographic information:
Hildur Jónsdóttir
kt.170377-3389, 2020, A Parallel Icelandic Dependency Treebank: Creation, Annotation and Evaluation, M.A. thesis, Faculty of Icelandic and Comparative Cultural Studies, University of Iceland.

Printing: Háskólaprent, Fálkagata 2, 107 Reykjavík
Reykjavik, Iceland, January 2020

Þessi ritgerð er tileinkuð dætrum mínum, Valgerði, Gunnhildi og Ásdísi

Abstract

This thesis describes the creation, annotation and evaluation of an Icelandic dependency treebank. This treebank holds syntactic annotation that is necessary for parser development and grammar research. Syntactic parsers are useful in various types of information technology applications and treebanks are the essential training data for data–driven natural language parsers. Parallel corpora have been mainly used for training machine translation systems but can also be used for creating dictionaries and ontologies, and multilingual and cross–lingual document classification. This first Icelandic parallel dependency treebank presented here is aligned with 19 other languages and is based on the Universal Dependencies (UD) annotation scheme. Studies on cross–lingual modeling have been growing constantly since the first UD treebanks were published and it could be a beneficial step for less-resource languages like Icelandic to become a part of this international research. Creating a treebank can be an extremely laborious task and it is therefore important to utilize accessible methods and data applicable for research. Here the method of preprocessing syntactic relations using delexicalized parsing was explored. The description of dependency grammar for Icelandic according to the UD annotation scheme is documented in appendix A and the Icelandic parallel UD corpus, Icelandic PUD, will be published as part of the UD project, version 2.6.

Útdráttur

Þessi ritgerð lýsir gerð, greiningu og mælingum á íslenskum samhliða venslatrjábanka. Þessi trjábanki inniheldur setningafræðilega greiningu sem er nauðsynleg bæði fyrir þróun á þáttara og setningafræðilegar rannsóknir. Þáttarar sem byggja á setningagreiningu eru gagnlegir í margvíslegar tegundir af textavinnslu í upplýsingatækni og eru aðal stoðgögnin fyrir þjálfun á gagnadrifnum þátturum. Samhliða málheildir hafa aðallega verið notaðar í þjálfun á þýðingarkerfum en einnig er hægt að nota þau í uppbyggingu á orðabókum og skjalaflokkun á ólíkum tungumálum. Þessi fyrsti samhliða venslatrjábanki fyrir íslensku sem er kynntur hér er til á 19 öðrum tungumálum og byggir á alþjóðaverkefninu Universal Dependencies (UD). Rannsóknir á þvermállegum líkönum hafa aukist verulega síðan UD verkefnið var sett á laggirnar og það er vænlegt skref fyrir tungumál eins og íslensku að verða hluti af alþjóðlegri rannsóknarvinnu. Mikil vinna felst í því að búa til nýjan trjábanka og þess vegna er mikilvægt að nýta sér aðgengilegar aðferðir og hagnýt gögn. Í þessari ritgerð prófa ég aðferð til að forvinna setningafræðileg vensl með því að nota aflexíkalíserað líkan. Lýsing á venslamálfræði fyrir íslensku samkvæmt UD greiningarskemanu er skjalað hér í viðauka A og nýi íslenski trjábankinn, Icelandic PUD, verður gefinn út með UD verkefninu, útgáfu 2.6.

Preface

My master's studies started in autumn 2003 when the second class of students in language technology were introduced to this new subject in Iceland. The first and the second class numbered around 10 students with different backgrounds, either linguistics or engineering. I finished my bachelor degree the same year with Latin as major and Icelandic as minor and before that I spent one year studying software engineering. From an early age, I studied music and foreign languages, along with the obligatory studies of Danish and English; French, German, Italian, Spanish, Old-French and Attic Greek. The language technology field seemed to combine all my interests. At that time, the language technology program was run in collaboration with the Nordic Graduate School of Language Technology and I had the opportunity to take a course on treebanks in Växjö, led by Joakim Nivre. The growing interest in treebanks and especially dependency grammar and parsing had begun in Växjö, the development of MaltParser (Nivre et al., 2007) and Joakims PhD on dependency parsing (Nivre, 2005b). For my assignments in the treebanks course, I needed syntactically annotated data but no Icelandic treebank had been published. Fortunately, an old friend, Gunnar Hrafn Hrafnbjargarson, was working on the Icelandic version of a Nordic parallel treebank and sent me a subset. From this time on, I've found treebanks fascinating and an obvious choice for my master's thesis. In the summer of 2006 I did an analysis on the need and design for an Icelandic treebank and started work on a thesis which had the title *Recognizing long-distance dependencies in Icelandic using data-driven approach*. Neither the analysis nor the thesis were ever finished. It wasn't until last year that I had the opportunity to revisit my studies and it is with gratitude that I hand in this thesis and the first parallel Icelandic dependency treebank.

Contents

List of Figures

List of Tables

Abbreviations

NLP – Natural Language Processing
UD – Universal Dependencies
PUD – Parallel Universal Dependencies
PoS – Part-of-Speech
TTR – Type–Token Ratio
MSP – Mean-Size of Paradigm
LD – Lexical Diversity
LAS – Labeled Accuracy Score
UAS – Unlabeled Accuracy Score

Acknowledgments

I would like to express my thanks of gratitude to Eiríkur Rögnvaldsson, Professor Emeritus, who has been the primus motor for Language Technology in Iceland. Eiríkur supported my primary work on treebanks in 2006 and encouraged me to finish my studies. It has been a valuable experience to study with the guidance of my advisor, Anton Karl Ingason, without his help this thesis would not have been possible. The treebank team at Árni Magnússon Institute for Icelandic Studies; Einar Freyr Sigurðsson, Research Lecturer, Þórunn Arnardóttir and Hinrik Hafsteinsson deserve special thanks in collaborating on the UD Annotation Scheme. It was a privilege to work at The Árni Magnússon Institute for Icelandic Studies and become acquainted with the team, especially Kristín Bjarnadóttir who was always there to discuss grammatical issues enthusiastically. Special thanks to Hrafn Loftsson, Associate Professor at Reykjavik University, who reviewed my draft on this work and Svanhvít Ingólfsdóttir who has also given me feedback and encouragement. I thank Hákon Sigurhansson, Director at Origo, for giving me the flexibility and support at work to focus on my studies and the translator, Ölvir Gíslason, for his understanding on the nature of the project. Finally, I thank my husband Steindór S. Guðmundsson, for a beautiful friendship, spirit of adventure and endless support.

1. Introduction

In order to survive the competition with a global language of 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 (UD) project (Nivre et al., 2016) has been a leading force in parsing and cross–lingual research and becoming a part of it could make Icelandic Language Technology more viable. An Icelandic treebank based on this type of an annotation scheme could also become a foundation for further Icelandic parser development. 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. A parser for Icelandic could for example support the 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 it was decided to start with studying the UD annotation scheme by working on a small corpus. The small corpus selected for the project is the Parallel Universal Dependencies (PUD) which has 1,000 sentences. This is 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).

As the core of the UD project is about consistency and parallelism the focus was 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 led by Einar Freyr Sigurðsson at The Árni Magnússon Institute for Icelandic Studies, carried out work on a conversion tool from the IcePaHC treebank (Rögnvaldsson et al., 2012a) to UD. The two groups collaborated on finding the best

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

solution 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 could be helpful in developing the Icelandic annotation scheme. My aim is to stay as unbiased as possible by previous work done in treebanking for Icelandic and related languages while utilizing tools, data and methods developed for the task.

The parallel corpora in UD are based on newspaper texts and Wikipedia data which is a genre not part of the corpus to be converted. The process for creating the Icelandic PUD is described in this paper. The raw source data was translated from the English source to Icelandic and then it went through automatic tagging, lemmatizing, conversion to CoNLL–U format, preprocessing the syntactic annotation with delexicialized methods and finally manual verification and UD validation. The Icelandic PUD will be published as part of UD version 2.6 [2].

1.1. Treebanks

Treebanks are text corpora annotated with syntactic or semantic structure. In this paper henceforth, the term *treebanks* refers to treebanks with syntactic annotation unless otherwise noted. Treebanks are the essential training data for data–driven natural language parsers and are also applied in cross–lingual learning and grammatical research. The development of treebanks started in the early 1990s and was a revolutionary step in computational linguistics research. The first large–scale treebank was published in 1993, The Penn Treebank (Marcus et al., 1993), with 4.5m tagged words in American English. The Penn Treebank project produced important corpora in the following years, Brown, Switchboard, The Wall Street Journal and more (Yurafsky and Martin, 2018). The annotation scheme is based on phrase–structure grammar and in later versions the use of traces and more details like grammatical function, text categories and semantic function were added among other improvements and amendments to the scheme. The first Icelandic treebank, IcePaHC (Rögnvaldsson et al., 2012a) is based on the Penn Historical Treebank project, see figure 1.1 using Penn PoS tags, phrase and function annotation and empty nodes which mark long–distance dependencies or syntactic movement.

With growing interest in adding functional annotation to treebanks, the development of dependency treebanks followed and was initiated for languages having relatively free word–order like Basque, Czech, German and Turkish (Kakkonen, 2006). Some treebanks have also been developed using both phrase–structure and dependency frameworks like Tiger (Brants et al., 2002) and Arboretum(Bick, 2003).

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

```
( (IP-MAT (NP-SBJ (N-N Sonur-sonur)
        (NP-POS (NP (NPR-G Glúms-glúmur))
            (CONJP (CONJ og-og)
                (NP (NPR-G Þórdísar-þórdís) (NPR-G Ásmundardóttur-ásmundardóttir)))))
    (BEDI var-vera)
      (NP-PRD (NPR-N Óspakur-óspakur)
          (, ,-,)
          (CP-REL (WNP-1 0)
              (C er-er)
              (IP-SUB (NP-SBJ *T*-1)
                  (VBDI deildi-deila)
                  (PP (P við-við)
                      (NP (NPR-A Odd-oddur) (NPR-A Ófeigsson-ófeigsson)))
                  (CP-ADV (WADVP-2 0)
                      (C sem-sem)
                      (IP-SUB (ADVP *T*-2)
                          (NP-SBJ *exp*)
                          (VBPI segir-segja)
                          (PP (P í-í)
                              (NP (NP-POS (NPRS-G Bandamanna-bandamaður))
                              (N-A sögu-saga))))))))
      (. .-.))
  (ID 1310.GRETTIR.NAR-SAG,.18))
```

Figure 1.1: Icelandic sentence in the Penn Historical Annotation Scheme

Later on, the conversion of treebanks from constituent to dependency framework was practiced and conversion tools developed (Penn2Malt[3] and LTH converter (Johansson and Nugues, 2007)). More recently the conversion of constituent and dependency treebanks has been focused on the UD framework specifically, e.g. the Stanford Dependencies and Penn treebank conversion to UD (Peng and Zeldes, 2018).

As a sign of the growing popularity of dependency treebanks and the success of the Universal Dependencies project, the last annual international workshop on Treebanks and Linguistic Theories (TLT) was held in cooperation with The International Conference on Dependency Linguistics (Depling), and the Universal Dependency Workshop (UDW) (Candito et al., 2019).

[3]https://cl.lingfil.uu.se/ nivre/research/Penn2Malt.html

1.2. Dependency and Phrase–Structure Annotation

Most syntactic treebanks are annotated with phrase–structure or dependency grammar structure. The main difference between dependency and phrase–structure grammar lies in the division of a clause where phrase–structure grammar splits the clause into a subject and predicate while dependency grammar considers the main verb as the root of the clause structure, see figures 1.2 and 1.3. The dependency grammar has flatter structure than phrase–structure grammar and is characterized by binary relations between words which are labeled with functional categories. However the phrase–structure groups words into phrases which are labeled with structural categories. The benefits of the dependency tree is that the relations give direct information on the structure where it is buried in the phrase–structure tree, e.g. the syntactic relation between *Suðurlandi* in the prepositional phrase and the main verb *hafði* goes through four levels in the phrase–structure whereas the relation is direct in the dependency structure. Also the dependency relations give approximation to the semantic relationship between predicates and arguments.

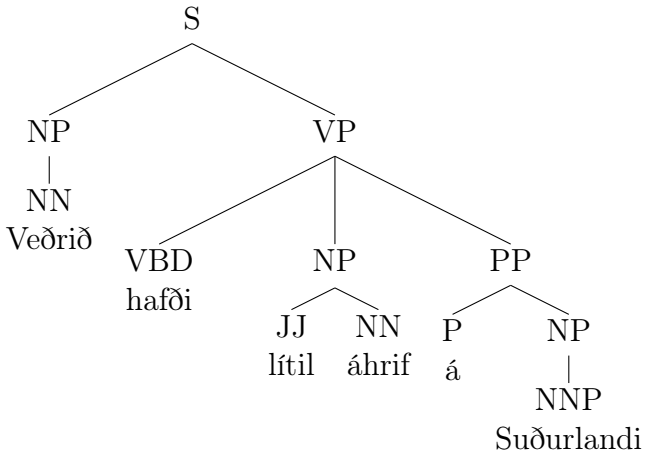


Figure 1.2: Phrase–structure for an Icelandic sentence

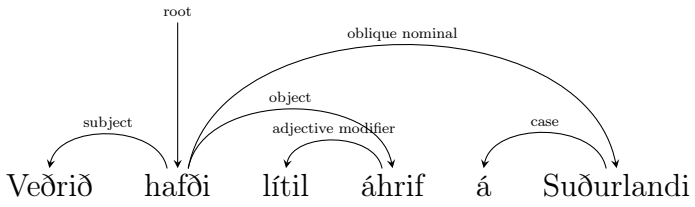


Figure 1.3: Dependency structure for an Icelandic sentence

It is important to acknowledge that these two frameworks are not opposites and there are hybrid frameworks that combine elements from both frameworks, e.g. Head–

Driven Phrase–Structure Grammar (HPSG). When certain elements are available it is often possible to convert the annotation between the two frameworks. In the Penn Historical Annotation Scheme the functional categories are included in the phrases so the conversion from IcePaHC to dependency grammar is well plausible and has undergone a proof of concept (Kárason, 2016).

1.3. Related work

1.3.1. Icelandic Corpora and Parsing Systems

The Icelandic Frequency Dictionary is the first text corpus used in natural language processing (NLP) in Iceland (Pind et al., 1991). It contains 500,000 tokens (henceforth, total number of words in a text) from 100 different texts, mostly taken from literary texts published in 1980 – 1989. These kinds of projects were originally led by two different professions, typists for their interest in the most frequent words in the language to increase their typing speed and secondly by educationists for their interest in teaching. At this time, programmers became interested in statistics of language with the mission of decreasing storage space and also for the development of automatic grammar checkers. Linguists were also becoming more interested in using computers in their research. The corpus was tagged with word classes and morphological features, preprocessed with automatic scripts and manually reviewed. The outcome was printed in many types of lists of frequencies and later made accessible on the internet. This groundbreaking work for Icelandic NLP became essential in Icelandic part–of–speech (PoS) tagger research (Helgadóttir, 2005). The IFD tagset[4] referred to in this project is a revised version of this original work.

With the growing demand on resources for NLP tasks, the first Icelandic treebank appeared in 2011 (Rögnvaldsson et al., 2012a). The parsing scheme was originally designed for the Penn Historical Corpora for Historical English and it uses phrase–structure annotation in a labeled 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. In 2019 the conversion of IcePaHC to UD started with the side project of adding modern text, about 100 k tokens, to the corpus. Another Icelandic treebank which is based on wide–coverage context free grammar is now being developed at Miðeind (Þorsteinsson et al., 2019). It is also planned to convert it to a dependency annotated treebank for training deep neural network–based parsers.

[4] `http://www.malfong.is/files/ot_tagset_files_en.pdf`

Another notable recently published corpus 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 it could also be used for e.g., creating dictionaries and ontologies, multilingual and cross–lingual document classification. This corpus is tagged with the IFD tagset, described in section 3.3. Research in machine translation reveals that leveraging syntactic knowledge using a dependency parser improves the translation quality (Wang et al., 2018).

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). The main goal of the Icelandic dependency corpora being built is to develop a dependency based parser for Icelandic to be applicable for NLP tasks.

1.3.2. Universal Dependencies for Related Languages

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. In this section, UD treebanks for related languages are described.

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 annotation guidelines for the original treebank were developed independently in an iterative process and based on the Norwegian Reference Grammar (Faarlund et al., 1997). The annotations were made with consideration to similar treebanks, the Swedish treebank Talbanken and the treebank of old Indo-European languages PROIEL. The NDT is divided into Bokmål (310 k tokens) and Nynorsk (301 k tokens) and contains mostly newspaper texts. A UD treebank of spoken dialects is available in Norwegian, LIA (Øvrelid et al., 2018), which was annotated with morphological and dependency–style syntactic analysis according to the LIA project and then converted to UD in 2017. The purpose of the corpus, which has 55 k tokens, is to increase research on spoken Norwegian with parser development in mind.

The Danish UD treebank (Johannsen et al., 2015) was converted from the Danish Dependency Treebank (DDT) (Kromann and Lynge, 2004) in 2015 . The DDT derives from a morphosyntactically tagged corpus created for a EU project called Parole (Bilgram and Keson, 1998) in 1996-1998. The linguistically annotated subcorpus of Parole holds 250,209 tokens in 16,062 sentences from 1,553 different texts. The texts are of various genres, mainly newspaper texts. DDT holds 100,200 tokens in 5,540 sentences from 536 randomly selected texts in Parole. The grammar is based on discontinuous grammar and the annotation guidelines are available in a 110 page documentation. The conversion of DDT to UD was implemented with a test-driven approach. A gold standard of 28 sentences was used as a quality reference for the conversion tool. A goal subset of 17 sentences including basic syntax was expected to achieve LAS of 100%. The other 11 sentences include rarer syntactic structures and the whole set was used to measure overall quality. The resulting LAS for the goal subset achieved 100% and the overall reference set scored 86.44% in LAS and 89.54% in UAS.

In version 2.5 of UD there are 3 different Swedish treebanks. UD-Talbanken has been a part of UD since version 1 and 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 texts, online manuals and Europarl data and number about 90,000 tokens in total. The third Swedish treebank, Swedish-PUD, has been a part of UD since version 2 and is a part of the Parallel Universal Dependencies (PUD). The texts are mainly from news and Wikipedia, translated to Swedish from English. The syntactic annotation is manually verified and the morphological annotation is automatic. Because of its size, it is available as a test file in UD like all PUD treebanks.

For Faroese, which is the closest relative of Icelandic and spoken by only 72,000 people, there is a UD corpus with 10 k tokens including texts from Faroese Wikipedia, released in November 2018. The Faroese UD treebank, although small, has been used as target language for two recent studies on cross-lingual dependency parsing with interesting results (Barry et al., 2019; Tyers et al., 2018).

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. The main goal of cross-lingual studies is to provide data and tools for low–resource languages and it is very pleasant to see the studies already done with the Faroese UD treebank which are promising and motivating for Icelandic UD research.

2. Dependency Grammar and Parsing

2.1. Dependency Grammar

2.1.1. History of Dependency Grammar

Dependency grammar is first mentioned in Panini's grammar of Sanskrit but the starting point for modern theoretical tradition is the work of Lucien Tesniére in the late 1950s (Keith Percival, 1990). Meanwhile, the concepts of dependency, governor, object and agent are used in other grammar descriptions, e.g., Arabic (Kiss and Alexiadou, 2015). Tesniére's book, ‚Elements de Syntaxe Structurale', was first published in 1959, 5 years after his death and did not become worthy noted until recently (Tesniere, 2015). He developed a method of structural analysis which is known as the dependency tree. Tesniére followed the trend of modern predicate-argument structure and thus deviated from the classic Aristotle subject-predicate division. This 700-page book attempts to describe the universality of human languages with syntactic analysis of about 5,000 sentences in over 60 languages. The concepts of valency and head-initial vs. head-final languages are considered to be the most valuable contribution (Tesniere, 2015).

The valency concept concerns the structure of verbs and its arguments which he connected to the semantics of the verb and seemed to be common across languages. This is what we recognize today as the predicate–argument structure, see figure 2.1.

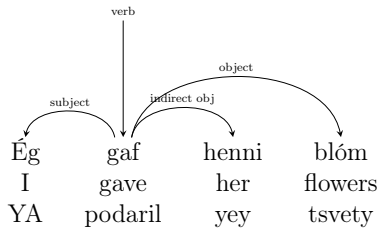


Figure 2.1: Ditransitive valency type in Icelandic, English and Russian

Tesniére also describes the concepts of head-inital and head-final languages to classify languages. Head-inital structure has the head of the structure appearing on the left and head-final structure has the head on the right, see figure 2.2. He classifies French languages as mitigated head-inital as it contains more head-initial than head-final structures and when measured in the French PUD treebank this is correct, see table 2.1. All dependency relations in these PUD treebanks are counted and classified in head-inital or head-final structure. The Germanic languages in PUD are also more head–initial than head–final and the Icelandic language is even closer to the middle.

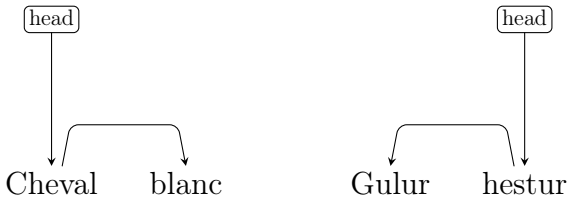


Figure 2.2: Examples of head-initial (left) and head-final (right) structures

Language	**Head-initial ratio**	**Head-final ratio**
French	59.42%	40.58%
Icelandic	54.20%	45.80%
Swedish	60.45%	39.55%
English	63.93%	36.07%
German	62.62%	37.38%
Czech	56.86%	43.14%

Table 2.1: Head-inital and head-final structures of 6 PUD treebanks

The aim is not to give details of all the aspects of Tesniére's studies here, an introduction to his work and an English translation are freely available[1] for further reading.

Since Tesniére, more theories for dependency grammar have been developed, including Prague's School Functional Generative Description (FGD), Melcuk's Meaning-Text Theory (MTT) and Hudson's Word Grammar (WG) (Kübler et al., 2009; Nivre, 2005a). Universal Dependencies is also a framework based on dependency grammar which is nowadays probably the most widespread dependency framework, discussed in section 2.3.

Tesniére's work has not gained the attention it deserves until recently. It has been stated that Chomsky's work, *Syntactic Structures*, published in 1957 and also the fact that the *Elements de Syntaxe Structurale* is written in French might have contributed to this (Tesniere, 2015).

[1] `https://www.jbe-platform.com/content/books/9789027269997`

Since the emergence of NLP, the focus on dependency grammar has grown to a great degree in recent years, mainly because of the ability to parse natural languages, described in section 2.2.

2.1.2. Basic Concepts of Dependency Grammar

The core of dependency grammar is the idea that words are linked by binary, asymmetrical relations called dependency relations. The dependency relation holds between the head and its dependent. The dependency relation can be labeled with a dependency type. An example of these dependency relations and types is given in figure 2.3 where syntactic dependency relations connect from the head to its dependent. The adjective *áhugasamur* is a dependent of the word *gítarleikari* and the verb *hélt* is the head for the object *tónleika*. This representation is according to UD guidelines. However the arrows are sometimes presented vice versa in the literature. Other conventions of representing dependency relations exist but they all have the same purpose of grouping syntactic units. To ease formal definitions and computations all words in the sentence must have a head so an artificial relation is created, called *root*, for the center of the clause which is most often the finite verb. In UD and this paper the terms *head* and *dependent* are used. However the terms modifier or child are sometimes used in the literature instead of dependent, and governor, regent or parent instead of head (Kübler et al., 2009). This paper is focused on syntactic dependencies. However the dependency relations can also be used for other types of relations like semantics, morphology (Polguère and Melcuk, 2009) or prosody (Groß, 2014).

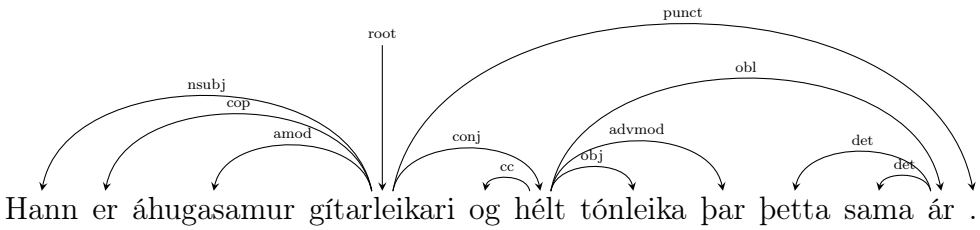


Figure 2.3: Dependency structure

Projectivity is a term in dependency grammar that captures the word order in a sentence. A dependency tree is projective if it has no crossing relations and the majority of sentences in Icelandic and related languages are projective. Non-projective trees are displayed in figure 2.4 where the root of the relative sentence, *kom* refers to the object of main sentence, *bók* but the temporal oblique *í fyrradag* interferes and causes the dependencies to cross. These structures are more frequent in languages that have flexible word order. The notion of projectivity is important because automatically

converted treebanks from phrase-structure to dependencies using head-finding rules are missing non-projectivity since it is retrieved from context-free grammar. Projectivity is also important in parsing because the non-projectivity relations cannot be parsed with the commonly used transition-based approaches (see section 2.2) but need more complex algorithms (Yurafsky and Martin, 2018).

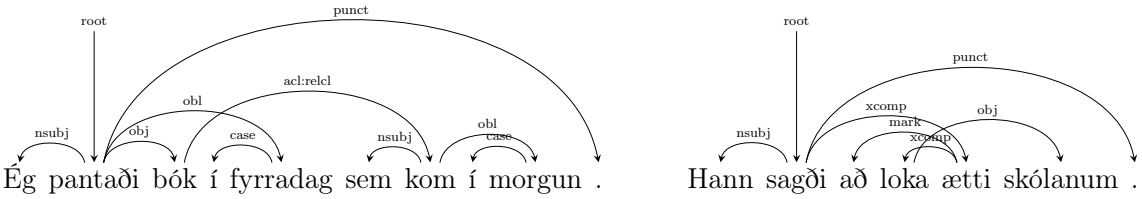


Figure 2.4: Two examples of non-projective sentences

2.2. Dependency Parsing

Syntactic parsers are useful in various types of information technology applications. Parsers automatically provide the structure of a sentence which leads us towards the meaning. In recent years, dependency-based methods for syntactic parsing have become increasingly popular for many reasons. One of them is that the dependency grammar has a clear predicate-argument structure that serves well in machine-translation and information extraction tasks and which is better suited for languages with flexible word order. Dependency treebanks have also been the input to the development of accurate syntactic parsers (Kübler et al., 2009). The time complexity of dependency algorithms is usually more appealing than for phrase-structure algorithms (Nivre et al., 2007) since dependency relations are binary but phrase-structure relations are one-to-one or many as described in section 1.2.

Since the beginning of dependency parsing, many algorithms, models and combined approaches have been proposed: Parsing with Dynamic Programming (Eisner, 1996), Graph algorithms MST (McDonald et al., 2005), Constraint Satisfaction (Karlsson, 1990) and Deterministic Parsing (Nivre, 2003). Algorithms worth mentioning include transition-based, graph-based and grammar-based models. Grammar–based models are either context-free or constraint-based and because of the increase in dependency treebanks the data–driven models, transition–based and graph–based, have been used more in research and parser development. Besides using treebanks as training data, pre–trained word embeddings are now more frequently applied in parser systems (Zeman et al., 2018).

The accuracy of the best dependency parsers for English reach over 95%[2] on unlabeled accuracy score (UAS) and labeled accuracy score (LAS) (see section 4) often with ensembled methods, like combining the strengths of constituencies and dependencies in HPSG parsing (Mrini et al., 2019).

Cross–lingual dependency parsers are highly beneficial for less–resourced languages that need annotated resources. Cross–lingual parsing is based on the idea of utilizing common elements in different languages to parse new languages that have no or very little training data, like Icelandic. With the UD project, cross–lingual parsing has grown as the project has the goal of providing NLP for all languages, not only the resource–rich.

The annual Conference on Computational Natural Language Learning (CoNLL) always hosts a shared task where participants evaluate and compare their NLP systems. The shared tasks in 2006 and 2007 on parsing research gave a promising outcome but had a disadvantage on the evaluation part which was lacking a gold standard for tokenization and tagging, and the annotation schemes between languages were incompatible. In 2017 and 2018 the shared task on CONLL was about multi–lingual dependency parsing using the consistent UD treebanks. Out of 24 evaluated systems in CoNLL 2018, the HIT–SCIR parser gave the best average LAS score, 75.82%, tested on total of 82 treebanks in 57 different languages (Zeman et al., 2018).

2.3. Universal Dependencies Framework

The motivation for the Universal Dependencies project is the need for a comparative evaluation in multilingual research on syntax and parsing. Many treebanks have been created over the past years but most of them are language specific and therefore have their unique annotation scheme or grammar. Because of this, comparison between tools and data has been strenuous. An attempt to standardize corpora is made with the Universal Dependencies (UD) framework. The UD is a framework for cross–linguistically consistent grammatical annotation (McDonald et al., 2013). The goal is to facilitate multilingual parser development, cross–lingual learning, and parsing research. The UD annotation scheme aims to identify similar constructions across languages while allowing language–specific extensions. The following examples from the parallel UD treebanks illustrate the similarities, even between far related languages, see figures 2.5 and 2.6. Main grammatical relations are often the same, here the nominal subject and two oblique nominals. The annotation properties consist of a part–of–speech tag based on Google universal tags (Petrov

[2] `http://nlpprogress.com/english/dependency_parsing.html`

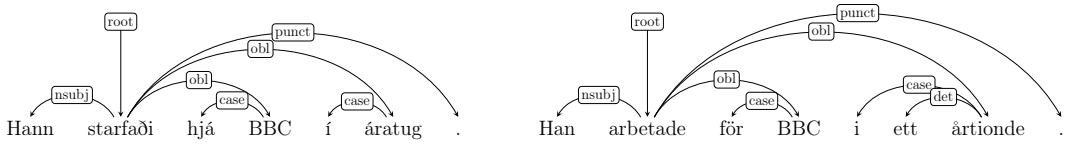


Figure 2.5: Parallel Icelandic and Swedish sentences from PUD treebanks

et al., 2012), morphological features from Interset interlingua (Zeman and Resnik, 2008) and syntactic dependencies, an evolution of Stanford Dependencies (Marneffe et al., 2014).

The syntactic annotation aims to capture the core arguments from other dependents in the sentence. The dependency relations are split into three structures: nominals, clauses and modifier words. The dependency relations used in the Icelandic annotation scheme are explained with examples in section A.3.

In 2015, 37 UD treebanks in 33 languages were released in UD and since then it has been constantly growing. The project has a "release–often" strategy with two releases a year and it is in continuous development. All contributions to this open community is welcome and the only restriction to languages that already have provided data, is that the language documentation must be shared between different treebanks.

It is stated on the UD website that the design has everything to do with the success of the project. The manifesto consists of 6 elements [3]:

1. UD needs to be satisfactory on linguistic analysis grounds for individual languages.

2. UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross–linguistic parallelism across languages and language families.

3. UD must be suitable for rapid, consistent annotation by a human annotator.

4. UD must be suitable for computer parsing with high accuracy.

5. UD must be easily comprehended and used by a non–linguist, whether a language learner or an engineer with prosaic needs for language processing. We refer to this as seeking a habitable design, and it leads us to favor traditional grammar notions and terminology.

6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, . . .).

[3]https://universaldependencies.org/introduction.html

In the latest release, UD version 2.5, there are 157 treebanks in 90 languages available. Without doubt, this can be considered a success and there are currently 18 treebanks in 16 languages upcoming, Icelandic being one of them.

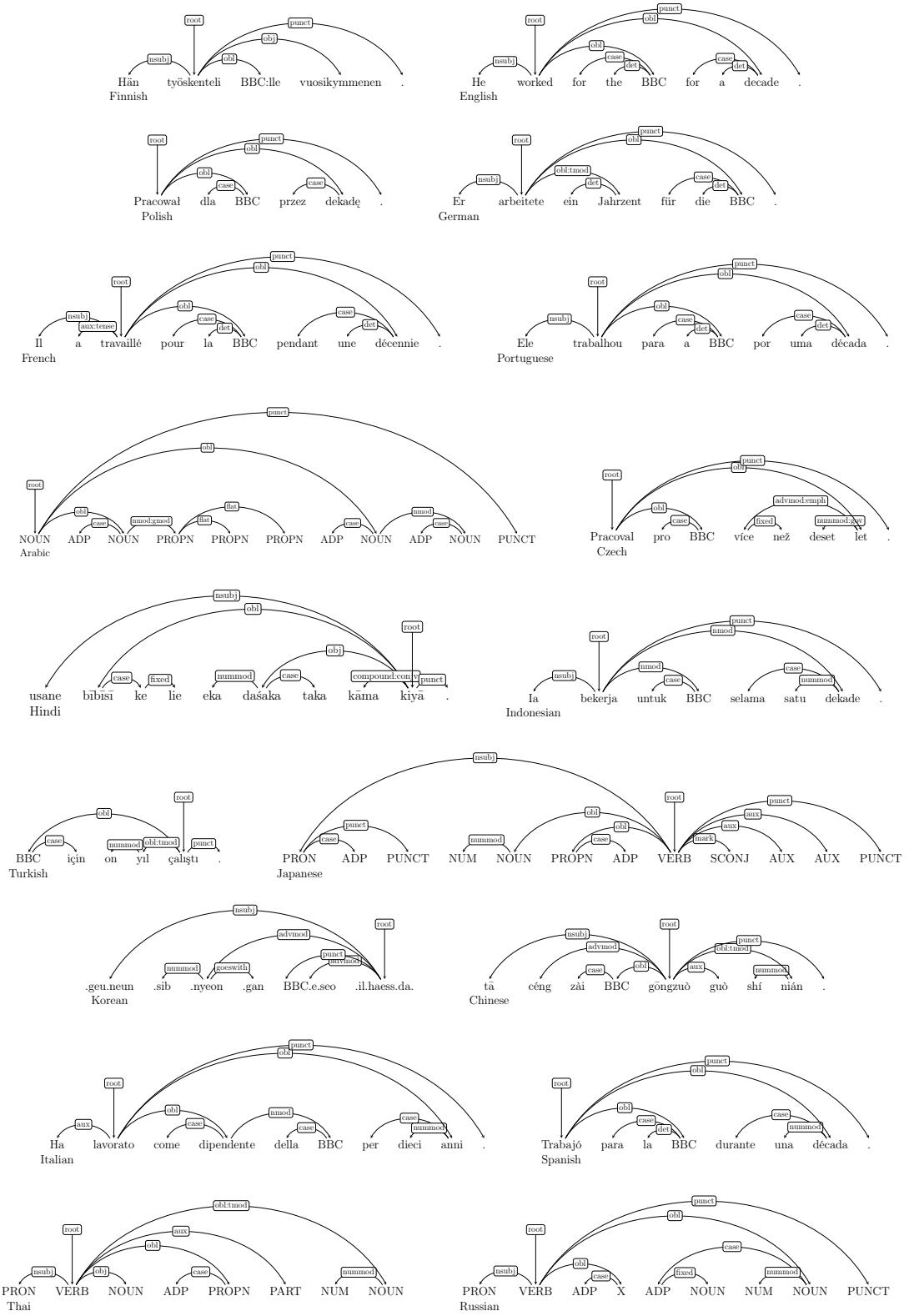


Figure 2.6: Parallel sentences from 18 PUD treebanks

3. Data, Annotation Scheme and Tools

3.1. Source Data

Along the conversion of IcePaHC to UD it was 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 the Parallel Universal Dependencies (PUD). 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 in the PUD treebanks 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 in English but the remaining 250 sentences are originally in German, French, Italian or Spanish and translated to English which is the source language for the parallel corpora. This genre is not a part of the converted IcePaHC corpus which mainly consists of literary texts. Media texts often suite better for NLP tasks because they are more standardized than other genres (Solberg et al., 2014).

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 relatively small it is better suited for testing and evaluation than training purposes.

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 to Icelandic hence a professional translator, Ölvir Gíslason, was recruited to translate all the 1,000 sentences. He was only given the guidelines to let the sentences match accurately 1–1 to match the previous PUD treebanks. The translation has not been altered in

any way and gave exactly 1,000 sentences and 18,812 tokens.

3.2. Adjusting Icelandic to the UD Annotation Scheme

The UD project requires each language to share its annotation specification with other treebanks of the same language in order to increase consistency and parallelism. The focus when adjusting Icelandic to the UD annotation scheme was on alignment with related languages without losing any elements. The UD annotation scheme is based on dependency relations between lexical items, i.e. words. Words are not segmented further into morphemes but morphological features are kept as properties of words. The morphological representation of a word consists of a lemma, a part–of–speech tag and morphological features. The syntax annotation is split into basic and enhanced dependencies. The basic dependencies include two properties which are obligatory for all UD treebanks, the head relation for each word and the dependency relation type (the 'Universal Dependencies'). The enhanced dependencies are intended to make some of the implicit relations between words more explicit, and augment some of the dependency labels to facilitate disambiguation of arguments and modifiers. This includes annotation like ellipsis, propagation of conjuncts, controlled/raised subjects, relative clauses and case information.

The UD project uses the CoNLL-U format for the annotated treebanks. It has 10 tab–separated fields for each word and an empty line between sentences. Comment lines begin with hash (#). The properties for each field is described in table 3.1.

	Field	**Description**
1	ID	Index of the FORM
2	FORM	The word or lexical item
3	LEMMA	Lemma of the word
4	UPOS	The Universal part–of–speech tag
5	XPOS	The language specific part-of-speech tag
6	FEATS	The morphological features
7	HEAD	The syntactic information (the ID of the HEAD of this word)
8	DEPREL	The dependency relation of the HEAD
9	DEPS	The enhanced dependency graph
10	MISC	Any other annotation

Table 3.1: Properties of the CoNLL–U format

The mandatory elements for a UD annotation scheme are the lexical item, the part–of–speech tag and the basic dependencies (i.e. ID, FORM, UPOS, HEAD and

DEPREL). All other fields are optional. Related languages include the lemmas and features and two treebanks, the Swedish Talbanken and Swedish PUD have some enhanced dependencies. Both the lemmas and most features are available in IcePaHC and other Icelandic corpora can retrieve those with automatic tools (see sections 3.3 and 3.4) so it was decided to include them as well for the Icelandic UD treebanks. The tagset designed for the IcePaHC corpus differs in some ways from the IFD tagset described in section 3.3 which is applied to most Icelandic corpora so additional tagging was done for the IcePaHC conversion.

The Icelandic annotation utilizes all the Universal part–of–speech tags (UPOS), listed in table 3.2. The list of features is mainly based on the IFD tagset and lemmas, see 3.5. All the main features are parallel with related languages however e.g. the Norwegian annotation scheme includes the feature *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 in the feature values as Icelandic has more inflections than related languages.

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 3.2: 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 3.3: List of lexical and inflectional features

The dependency relation types chosen for the Icelandic annotation scheme are listed in table 3.4. In general the IcePaHC treebank holds very detailed syntactic in-

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

Table 3.4: Dependency relations for the Icelandic annotation scheme

formation so it was feasible to convert it to UD. Foreign names, brands, symbols and copula sentences were more noticeable in the Icelandic PUD whereas first or second person sentences and discourse elements were more frequent in IcePaHC. There are small variations in the 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. A description of the dependency relation types with Icelandic examples are listed in appendix A.3. Enhanced dependencies were not included in this first version but might be added later, e.g. the case information and the ellipsis which are a part of the IcePaHC annotation.

3.3. Tokenizing and Tagging with ABLTagger

For the properties of the UPOS and FEATS, the Icelandic translation had first to be tagged. The state–of–the–art ABLTagger was used which is based on BiLSTM models, morphological lexicon and a 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[1]. 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. The mapping from the IFD tagset to UPOS and features is explained in section 3.5 and appendix B. The ABLTagger also tokenizes the text utilizing a tokenizer from Miðeind[2] (Þorsteinsson et al., 2019) which greedily recognizes certain multi–token spans like dates and adverbial multi–word idioms.

3.4. Lemmatizing with Nefnir

Lemmas are not a mandatory property in the UD framework but for morphologically rich languages like Icelandic, the lemmas can be an important step in NLP tasks. For instance, in this project the lemmas were a key input in the mapping process, in particular for recognizing auxiliaries from other verbs and coordinating from subordinating conjunctions. Other NLP tasks that depend on lemmatization are machine translation, text mining and information retrieval. For lemmatization, 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 Inflections (Bjarnadóttir et al., 2019) to retrieve the correct lemma. It reaches an accuracy of 99.55% with verified tagged input, and for text tagged with a PoS tagger, the accuracy obtained is 96.88%. The outcome was converted as is, to the LEMMA property in the CoNLL–U format.

[1] `http://malfong.is/`
[2] `https://github.com/mideind/Tokenizer`

ID	FORM	LEMMA	UPOS	XPOS	FEATS
1	Rúmlega	rúmlega	ADV	aa	_
2	5,7	5,7	NUM	ta	NumType=Card
3	milljónir	milljón	NOUN	nvfn	Case=Nom|Definite=Ind|Gender=Fem|Number=Plur
4	Flórídabúa	Flórídabúi	PROPN	nkfe-s	Case=Gen|Gender=Masc|Number=Plur
5	hafa	hafa	AUX	sfg3fn	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin
6	þegar	þegar	ADV	aa	_
7	greitt	greiða	VERB	ssg	VerbForm=Sup|Voice=Act
8	atkvæði	atkvæði	NOUN	nhfo	Case=Acc|Definite=Ind|Gender=Neut|Number=Plur
9	í	í	ADP	aþ	_
10	tveggja	tveir	NUM	tfvfe	Case=Gen|Gender=Fem|Number=Plur
11	vikna	vika	NOUN	nvfe	Case=Gen|Definite=Ind|Gender=Fem|Number=Plur
12	utankjörfundarkosningu	utankjörfundarkosning	NOUN	nveþ	Case=Dat|Definite=Ind|Gender=Fem|Number=Sing
13	.	.	PUNCT	.	_

Table 3.5: Icelandic dependency annotation in CoNLL–U format

3.5. Mapping of IFD tags and Lemmas to UPOS and Features

The mapping 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*, *munu* and *skulu* (en. be, will, 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 the distinction between coordinating from subordinating conjunctions was made using the lemmas. All indefinite, demonstrative, interrogative and possessive pronouns are tagged with the a 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. However information on the pronoun is kept with the *PronType* feature. To hold parallelism to related languages the participles, both past and present, are mapped to UPOS adjective tag but the features hold information on the verb participle and therefore no information is lost.

An example from the Icelandic PUD in CoNLL–U format is given in figure 3.5. The details of the mapping from IFD to UPOS and morphological features can be found in appendix B.

3.6. Preprocessing Syntactic Relations with Delexicalized Modeling

Since no Icelandic dependency parser was available, a delexicalized parser was trained to preprocess the corpus. Delexicalized parsing, which is one type of cross–lingual model transfer, was first introduced in 2008 (Zeman and Resnik, 2008) and is currently considered a standard technique in cross–lingual parsing. Treebanks are available for only ∼1% of languages in the world and there are around 7.000 living

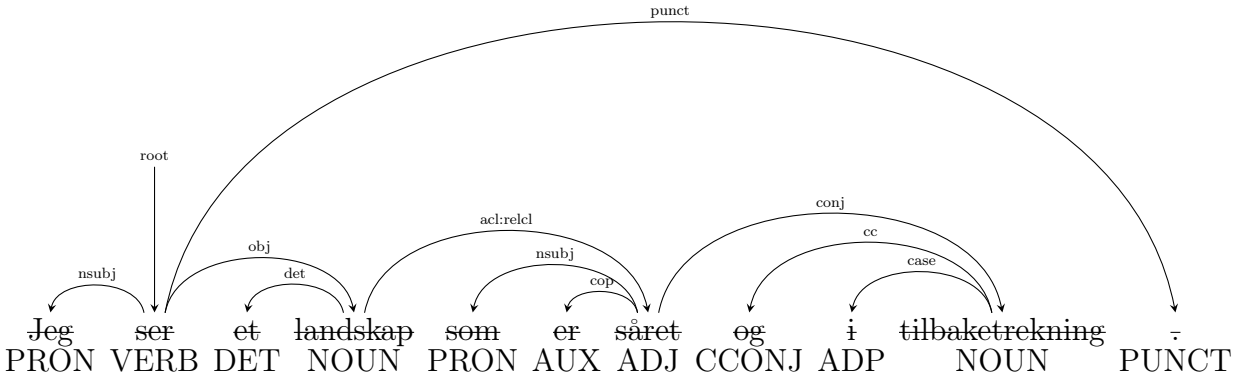


Figure 3.1: Norwegian training data for a delexicalized model

languages today[3]. One major goal of cross–lingual modeling is to provide robust NLP for all the targets, or the remaining ∼99%.

One experiment with cross–lingual parsing using Icelandic has been done with Faroese as target. A parser was trained with Faroese and Icelandic treebanks and the results showed that the accuracy of parsing Faroese increased with adding more trees from the Icelandic treebank, IcePaHC (Ingason et al., 2014).

Delexicalized models and annotation projections have been used to develop resources and tools for low–resource languages (Ganchev and Das, 2013). Other strategies have been shown to give good results for dependency parsing, e.g., machine translation with synthetic training data using parallel corpora (Tiedemann et al., 2014). The annotation projection would have been an interesting method to explore as well, however the development pipeline would need more programming and research.

The downside of cross–lingual modeling and annotation projection is that the manual reviewers may develop a bias from the annotation designed for the source language. To minimize the bias by manually reviewing automatic annotation, the Icelandic annotation scheme was first developed with documentation and examples.

Delexicalized models using only UPOS tags were trained with UD treebanks of related languages, Swedish, Norwegian, Danish and Faroese 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 format), see table 3.6.

The parser selected for the task is UDPipe (Straka and Straková, 2017) which was

[3] `https://www.ethnologue.com/`

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 3.6: Evaluation of delexicalized models

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. UDpipe implements Parsito parser which is a transition–based parser using a neural–network classifier. To include the non–projective transitions, the transition system "swap" was used which is a fully non–projective system and extends the projective system by 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.

Interestingly, the Swedish PUD model gave the second best results 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 19 k tokens, see table 3.6 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 it was 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 trained 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 a 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 sentences reached 78.82% UAS and 73.78% LAS. Figure 3.2 is an example of a sentence perfectly parsed by the last training model.

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

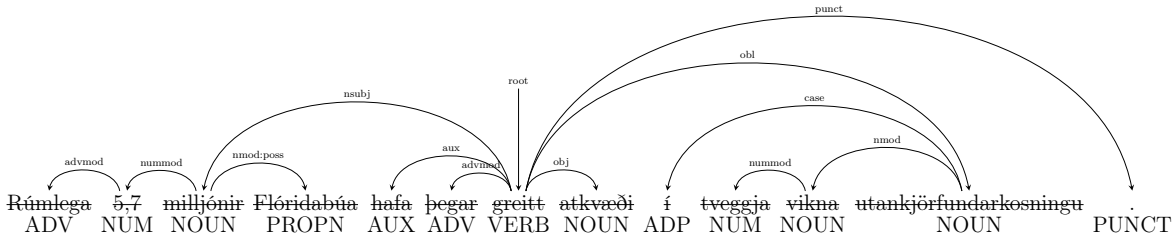


Figure 3.2: Correct parsing outcome using a delexicialized model

3.7. Manual Reviewing Process

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 different languages. The manual correction was done with UD Annotatrix (Tyers et al., 2017) that provides good graphical user interface for viewing and editing the annotation in graphical and CoNLL–U format. 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.

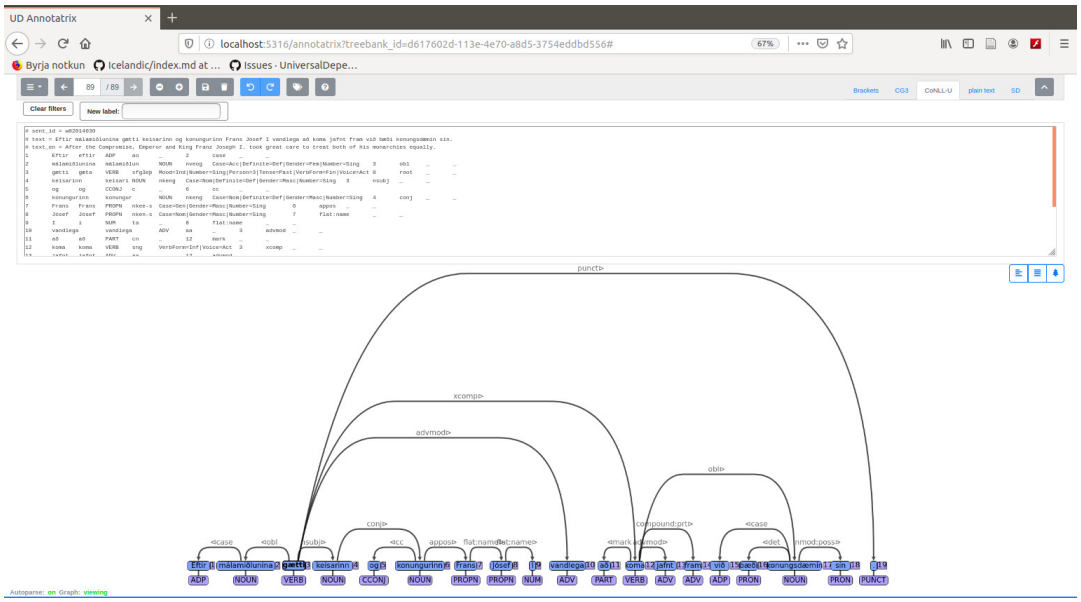


Figure 3.3: Screenshot from Annotatrix

4. Evaluation

4.1. Metrics

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 which is simply measuring the accuracy with a 10–fold cross validation or when the treebank is bigger than 100 K tokens, a 8–1–1 split to training, development and testing sets is sufficient[1]. To measure the quality of the Icelandic PUD a 10–fold cross validation will be applied using three different parsing models, see section 4.

When testing the parsability of a corpus using 10–fold cross validation, the corpus is divided into 10 complementary sets, each with 1 validation file (*gold file*) having 10% of the texts and 1 training file including 90% of the text. Each set is evaluated by training the parser with the training file and then the validation file is parsed with the trained model (outcome is the *test file*). The test file is then evaluated against the gold file. This is done 10 times and the average gives the final 10–fold cross validation result.

Many tools and scripts are provided by the UD project and one of them is an evaluation script to compare test and gold corpora. The validation script gives amongst other measures, information on tokens, sentences, words, UAS and LAS. For this project the major goal is to measure the UAS, unlabeled accuracy score and LAS, the labeled accuracy score since the focus is on the dependency relations.

The unlabeled attachment score (UAS) gives the percentage of words that have correct head and the labeled attachment score (LAS) gives the percentage of words that have correct head plus correct labels. Figure 4.1 gives an example of a sentence that has 6 words where a parsing model has 5 correct heads and 4 correct labels (right sentence). The UAS for this sentence is 5/6 and the LAS is (5/6 + 4/6)/2.

Metrics that draw out differences between languages are especially interesting when

[1] `https://universaldependencies.org/release_checklist.html#data-split`

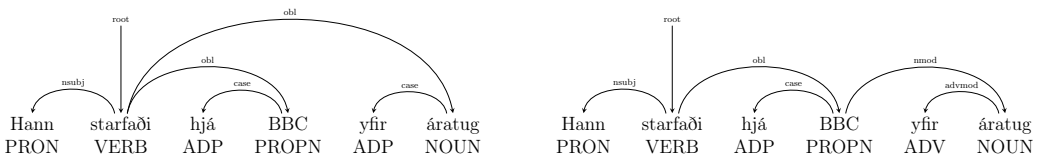


Figure 4.1: Gold sentence (left) and Test sentence (right)

working with parallel corpora. A standard way of measuring morphological richness is the type–token ratio (TTR) which is simply the tokens (total number of words) divided by token types (total number of distinct words). High value of TTR indicates rich morphology. Another metric is the mean–size of paradigm (MSP) which measures the number of token types per lemma, i.e., the number of lemma types (number of distinct lemmas) divided by the number of token types (Berdicevskis et al., 2018). Lexical diversity is also measured, the ratio of lemma types and total tokens. Other interesting metrics that can be considered are on word order rigidity and variability of UPOS tag sequences. However for reliable results the size of the corpora needs to be bigger so these were not measured.

4.2. Results

The UDPipe parser was chosen to evaluate the Icelandic PUD, the same one as used for preprocessing, see section 3.6. The default settings for the parser training model is using the FORM as the main dimension, this is called Model 1. The dimension refers to the size of the embedding the system uses for training. For Model 1, a dimension of 50 for the FORM and a dimension of 20 for UPOS, FEATS and DEPREL means that the model has more weight on the FORM, i.e., the word, than the grammatical annotation (UPOS, FEATS and DEPREL). The parser model was configured in two other ways. The second model is the simplest one, using only the UPOS and FEATS, this is Model 2. In the third model, the UPOS, FEATS and the LEMMA are used because the language is morphologically rich and the corpus very small so the lemma might be more suitable than the FORM. Table 4.1 shows the dimensions chosen for training the models.

The results show that the three models give very similar results. The lexicon information (FORM and LEMMA) does not increase the accuracy significantly which can be explained by the size of the treebank. If these results are compared to English, the default model (Model 1) was tested with 10–fold cross validation on the English PUD which gave 80.88% on LAS and 83.22% on UAS but for morphologically rich language like Czech the same model and evaluation gives 75.52% for LAS and 80.45% for UAS. It is not surprising that the accuracy is higher for the English

Dimensions	Model 1	Model 2	Model 3
UPOS	20	20	20
FEATS	20	20	20
XPOS	0	0	0
FORM	50	0	0
LEMMA	0	0	20
DEPREL	20	20	20

Table 4.1: Configuration for parsing models

	UAS	LAS
Model 1	79.415%	74.447%
Model 2	78.907%	73.762%
Model 3	79.920%	74.848%

Table 4.2: 10–fold cross validation results

PUD treebank as the Icelandic language is morphologically rich. The Czech language is not as related to Icelandic but was tested 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. The accuracy for the Czech PUD treebanks is slightly higher ($\tilde{1}$%) than for the Icelandic PUD. The vocabulary and morphological richness was also measured, see table 4.3.

	Icelandic	English	Swedish	Czech
Tokens	18,827	21,183	19,083	18,655
Token Types	6,756	5,734	6,402	7,896
Lemma Types	4,840	4,692	5,032	5,317
Mean-Size of Paradigm	1,396	1,222	1,272	1,485
Type–Token Ratio	35.88%	27.07%	33.55%	42.33%
Lexical Diversity	25.71%	22.15%	26.37%	28.50%

Table 4.3: Morphological complexity and lexical diversity of 4 PUD treebanks

The morphological complexity and lexical diversity were compared to the related PUD treebanks (Nivre et al., 2017), Swedish and English, and the Czech treebank is also measured because Czech is considered to be morphologically rich, like Icelandic. From the results in table 4.3 it is evident from values of TTR and MSP (see section 4.1) that Icelandic is morphologically richer than related languages. A higher ratio on the TTR was expected in the Icelandic corpus, compared to Swedish, but the lexical diversity of the Swedish data is higher which increases its TTR. These measures reveal the challenges in comparing languages, even with parallel data, rather

standardized text genre and accurate 1–1 sentence alignment. For example, with high lexical diversity which may vary between translators, the complexity measures (MSP and TTR) rise as well. Bigger datasets would provide more reliable data and opportunities for other measures like word order rigidity and variability of UPOS tag sequences. Hopefully the PUD data sets will be extended in the near future.

5. Conclusion

The first parallel treebank for Icelandic based on UD, Icelandic PUD, has been described. As a first step in adapting the dependency grammar with UD annotation scheme to Icelandic, 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 less–resource languages considering to participate in the UD project I believe that the source data and method described here is simple and convenient as a first step towards UD. In this case the work was valuable 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 ABLtagger with the IFD tagset in order to add the features.

Before annotating the Icelandic PUD corpus, some annotation experiments were done with Icelandic literature. One of the things that were expected is the number of non–projective structures (crossing relations, see section 2.1.2) that are more evident in languages with free or flexible word order like Latin (Straka et al., 2015). The Icelandic language is morphologically rich and offers relatively free word order. However in this work with the Icelandic PUD corpus it was, on the contrary, noticeable how rarely those constructions appeared with only 11 attested examples of such structures. Newspaper texts differ from other genres in that they are more standardized and therefore might be better suited for NLP tasks. Therefore the lack of non–projective structures could be explained by the text genre in this corpus which is informative and complex constructions are maybe not as frequent.

Regarding future work, the main goal of building treebanks is to build a parser for Icelandic. The converted Icelandic UD treebanks which are under construction (see section 1.3.1) will most likely provide sufficient data for training a parser model. Further work on a system based on the predicate-argument structure which has been developed for Icelandic (Rúnarsson, 2017) could be used to improve the dependency treebanks. This system called *Samba* identifies verbal expressions by using a database of verbal expressions in Icelandic, extracted from dictionary data and edited for syntactic analysis and machine readability (Bjarnadóttir, 2016). The

database of verbal expressions could also be extended with predicate–argument structure extraction from the Icelandic PUD and other Icelandic UD treebanks.

The UD framework is in continuous development and there are plans for two releases a year for new treebanks and updates or improvements of existing treebanks. There is always room for improvements in annotated corpora and I hope that users of the Icelandic PUD corpus will report issues for correction. Adding more features and sub–relations to the Icelandic annotation scheme and even enhanced dependencies like empty nodes for gapping and shared dependents in coordination would make the Icelandic UD treebanks even more interesting in the future.

I hope 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.

A. Documentation of Universal Dependencies for Icelandic

This documentation is based on version 2 of Universal Dependencies guidelines[1]. The UD project requires each language to share their annotation specification with other treebanks of the same language to increase consistency and parallelism. The focus when adjusting Icelandic to the UD annotation scheme was in alignment with related languages without losing any elements.

A.1. Morphology

A.1.1. Tags

A.1.1.1. List of POS tags

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 A.1: List of UPOS tags

The Icelandic annotation utilizes all the Universal part–of–speech tags (UPOS), listed in table A.1 like the Norwegian and Danish UD treebanks. The Swedish and Faroese treebanks are not utilizing the 'X' tag (other or unknown tag) and the

[1] `https://universaldependencies.org/guidelines.html`

Faroese treebank does not include the auxiliaries tag, 'AUX', nor the symbol tag, 'SYM'. The Faroese treebank indicates the auxiliaries using the syntactic annotation, see A.3.

The UPOS tags were automatically retrieved from the IFD tagset, see mapping in section B. Some tags required special attention and/or manual revision and are described here. The only word tagged with the tag PART is the infinitive marker *að*. Auxiliaries are all verbal in Icelandic and can be grouped into four types:

- The copula *vera* (be).
- The temporal auxiliary *hafa* (have) which combines with the supine form of the main verb to form perfect tenses.
- The passive auxiliary *vera* and *verða* (become) which combines with the past participle of the main verb to form passives.
- Modal and aspectual verbs that combine with the bare infinitive of the main verb, such as *mega* (may), *vilja* (want), *munu* (will) and *skulu* (shall) or the past participle of the main verb like *geta* (can) and *fá* (can) .

Copula is a connecting word, in particular a form of the verb 'to be', connecting a subject and complement. The copula term is not used in the traditional Icelandic grammar but the verb *vera* was automatically mapped to the tag AUX in the process and then manually corrected in the syntactic annotation, see section A.2.2.4.2. In the annotation, auxiliaries are only used for verbs that express grammatical category. That is, verbs that take compliments starting with the participle 'að' plus the infinitive are not tagged as auxiliaries.

The tag DET is used for articles and pronominal words used with a determiner function, including possessive, demonstrative and indefinite pronouns. The tag PRON is reserved for pronouns occurring as the head of a noun phrase.

A.1.2. Features

A.1.2.1. List of Features

The lexical and inflectional features (FEATS) chosen for the Icelandic annotation are listed in table 3.3, the strikethrough features were not included. All the main features are parallel with related languages. However e.g. 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 Icelandic has more inflections than the other North Germanic languages. As for UPOS in the last section, the FEATS are extracted from the IFD tagset, see B for details.

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 A.2: List of lexical and inflectional features

A.2. Dependency Relations

Syntactic information in the UD framework is presented by CoNLL–U format with field HEAD and DEPREL. The HEAD (field number 7))is the index of the head of the word, DEPREL (field number 8) is the label for the dependency relation. [2]

Table A.3 lists the dependency relations for the Icelandic annotation. In this chapter a description of the Icelandic dependency relations will be explained with examples. If the examples are taken from the Icelandic PUD corpus, the reference ID, the Icelandic translation and the original English version is provided.

[2] `https://universaldependencies.org/format.html`

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

Table A.3: Dependency relations for the Icelandic annotation scheme

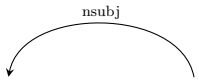
A.2.1. Core Arguments

A.2.1.1. Nominals

A.2.1.1.1 nsubj

The *nsubj* is the nominal subject of a phrase. For passive constructions, i.e. when the subject does not act as the agent of the phrase, it is recommended to add the subtype *nsubj:pass*. This subtype was not a part of the first Icelandic version as this annotation would need more research to be done consistently.

```
# sent_id = n01002042
# text = Þessi nýju útgjöld eru fjármögnuð með digrum bankareikningi Clintons.
# text_en = The new spending is fueled by Clinton's large bank account.
```



Þessi nýju útgjöld eru fjármögnuð með digrum bankareikningi Clintons .

Figure A.1: Dependency relation: nominal subject

A.2.1.1.2 obj

The object of a verb is a core argument of a verb with the subject. Typically, it is the noun phrase that denotes the entity acted upon or which undergoes a change of state or motion. In Icelandic the object is inflected in the accusative, dative or (rarely) genitive case.

```
# sent_id = w01051032
# text = Í ákafa sínum við að fanga Kadesh gerði Ramses II stórfelld hernaðarmistök.
# text_en = In his haste to capture Kadesh, Ramesses II committed a major
tactical error.
```

Í ákafa sínum við að fanga Kadesh gerði Ramses II stórfelld hernaðarmistök .

Figure A.2: Dependency relation: object

A.2.1.1.3 iobj

The indirect object of a verb is a core argument of a verb. Typically, it is the recipient of ditransitive verbs of exchange. In Icelandic the indirect object is inflected with the dative case.

```
sent_id = n01034060
text = Trudeau mun bjóða 45. forseta Bandaríkjanna þetta, sama hver hann eða
hún verður.
text_en = Trudeau will extend that invitation to the 45th president of
the United States, whoever he or she may be.
```

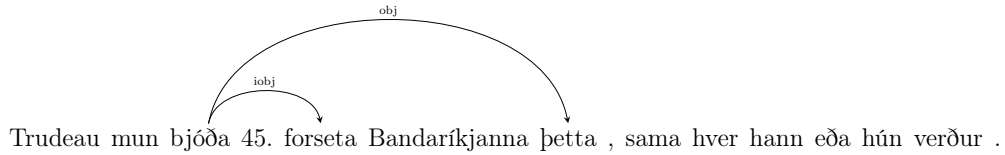


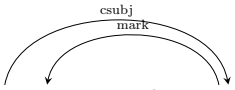
Figure A.3: Dependency relation: indirect object

A.2.1.2. Clauses

A.2.1.2.1 csubj

A clausal subject is a clausal syntactic subject of a clause, i.e., the subject is itself a clause. The head is the root of the main clause and the dependent is the root of the clausal subject.

```
# sent_id = n01101007
# text = Í eiðinum felst að löggjafar sverja Hong Kong, sem hluta af
Alþýðulýðveldinu Kína, hollustu sína.
# text_en = The oath involves lawmakers swearing allegiance to Hong Kong as part
of the People's Republic of China.
```



Í eiðinum felst að löggjafar sverja Hong Kong , sem hluta af Alþýðulýðveldinu Kína , ...

Figure A.4: Dependency relation: clausal subject

A.2.1.2.2 ccomp

A clausal complement is a dependent clause and a core argument. The head is the root of the main clause and the dependent is the root of the dependent clause.

```
# sent_id = n01087018
# text = Ég held að þess vegna hafi þau sökkt sér í mynstur og liti.
# text_en = I think that's why they immersed themselves in pattern and colour.
```

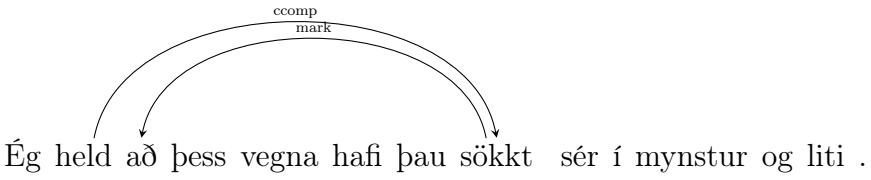


Figure A.5: Dependency relation: clausal complement

A.2.1.2.3 xcomp

xcomp stands for an open clausal complement or predicative without a subject. The head is the root of the main clause and the dependent is the root of the dependent clause.

```
# sent_id = n01046057
# text = Hann benti líka á að Rogers kynnti nýlega nýtt forritaverkfæri sem hjálpar
viðskiptavinum að fylgjast með gagnanotkun sinni.
# text_en = He also pointed out that Rogers recently introduced a new app tool that
helps customers monitor their data usage.
```

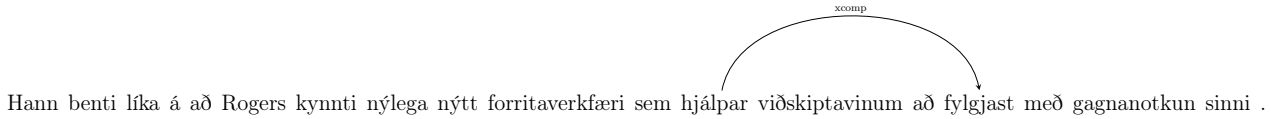


Figure A.6: Dependency relation: open clausal complement

```
# sent_id = n01120010
# text = Saga Doss er líka ólíkindaleg, sem gerir hana þeim mun meira hrífandi.
# text_en = Doss's story also has an unlikely quality to it that makes it all the
more appealing.
```

Figure A.7: Dependency relation: open clausal complement as a secondary predicate

A.2.2. Non–Core Dependents

A.2.2.1. Nominals

A.2.2.1.1 obl

obl is a non–core or oblique argument, usually a nominal. This is often a preposition phrase, dependent on the root of the sentence, or a bare nominal in inflected case.

```
# sent_id = n01127130
# text = Í hæstarétti Hong Kong, tveimur árum síðar, virtist Jutting vera
að mestu leyti yfirvegaður.
# text_en = In Hong Kong's High Court, two years later, Jutting appeared
largely composed.
```

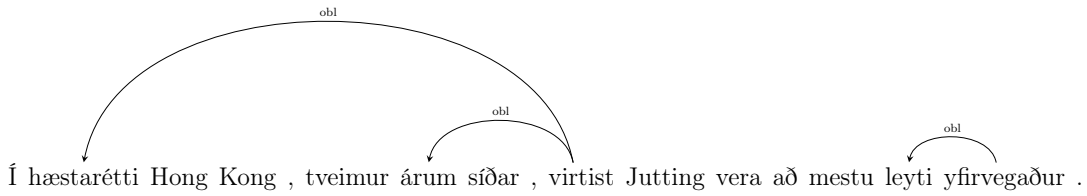


Figure A.8: Dependency relation: oblique dependent I

```
# sent_id = n01046003
# text = Farsímarnir okkar eru svo miklu meira en símar þessa dagana.
# text_en = Our cellphones are so much more than phones these days.
```

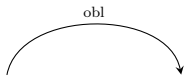


Figure A.9: Dependency relation: oblique dependent II

A.2.2.1.2 obl:arg

obl:arg, or oblique argument, distinguishes oblique arguments from adjuncts. It was introduced in version 2 of UD to identify prepositions that function as arguments (Zeman, 2017).

```
# sent_id = n01140012
# text = Í Hollandi beita yfirvöld ekki jafn tæknilegum aðferðum við að fylgjast
með drónum.
# text_en = In the Netherlands, authorities have taken a lower--tech approach to
tracking drones.
```

Figure A.10: Dependency relation: oblique argument

A.2.2.1.3 vocative

The vocative relation identifies a dialogue participant addressed in a text. In Icelandic, some words can be inflected in vocative case, e.g. *Jesú, vin, son*, i.e.'Jesus, friend, son', other words take nominative case when the function is vocative. As in Norwegian, adjectives take definite inflection when in vocative.

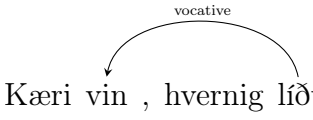


Figure A.11: Dependency relation: vocative

A.2.2.1.4 expletive

Expletives are challenging in syntactic annotation because they function as core arguments syntactically but not semantically. This label has received a special attention as its usage does not seem to be consistent within UD treebanks (Bouma et al., 2018). In the Icelandic PUD project the *expl* type is used for extraposition of clausal arguments A.12, existential A.13 and impersonal constructions A.14 and A.15, which aligns with the annotation in IcePaHC. It could be interesting for Icelandic grammar research to distinguish between the reflexive passive and the new passive so the passive reflexives might be taken into consideration for future versions. The inherent reflexives are all labeled as objects in IcePaHC and also in its conversion to UD. Grammatical details on Icelandic expletives are explained in (Þráinsson, 2007):312.

Figure A.12: Dependency relation: expletive in extraposition of a clausal argument

Figure A.13: Dependency relation: expletive in existential construction

Figure A.14: Dependency relation: expletive in impersonal construction (weather)

Figure A.15: Dependency relation: expletive in impersonal construction

A.2.2.2. Clauses

A.2.2.2.1 advcl

An adverbial clause modifier is a clause which modifies a verb or other predicate. This includes things such as a temporal clause, consequence, conditional clause, purpose clause, etc. The dependent must be clausal (or else it is an advmod) and the dependent is the main predicate of the clause.

```
# sent_id = w01016028
# text = Andrúmsloftið er óreiðukennt kerfi svo smávægilegar breytingar
á einum hluta þess geta haft mikil áhrif á kerfið í heild.
# text_en = The atmosphere is a chaotic system, so small changes to one
part of the system can grow to have large effects on the system as a whole.
```

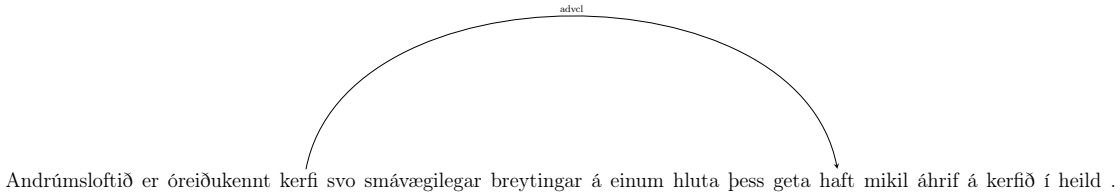


Figure A.16: Dependency relation: adverbial clause modifier

A.2.2.3. Modifier Words

A.2.2.3.1 advmod

An adverbial modifier of a word is a (non–clausal) adverb or adverbial phrase that serves to modify a predicate or a modifier word.

```
# sent_id = n01144041
# text = Ég var bara strákur í forugum skóm.
# text_en = I was just a boy with muddy shoes.
```

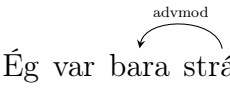


Figure A.17: Dependency relation: adverbial modifier

A.2.2.3.2 discourse

discourse is used for interjections and other discourse particles and elements. This relation never occurs in the Icelandic PUD.

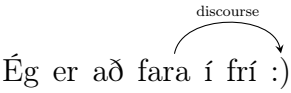


Figure A.18: Dependency relation: discourse

A.2.2.4. Function Words

A.2.2.4.1 aux

aux or auxiliary is a function word associated with a verbal predicate that expresses categories such as tense, mood, aspect, voice or evidentiality. When more than one auxiliary is a part of a verb predicate, the main verb is the head for all the auxiliaries which function as dependents.

```
# sent_id = n01148035
# text = Ekki hafa allar umbreytingar á svæðinu borið árangur.
# text_en = Not all transformations in the region have been successful.
```

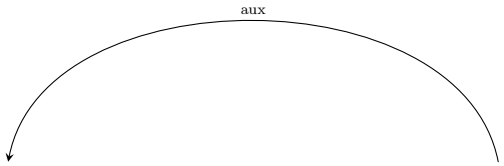


Figure A.19: Dependency relation: auxiliary

```
# sent_id = n02083054
# text = Það kemur ekkert á óvart að alríkisstjórnin og stjórnvöld ríkjanna skuli
hafa flokkað National Natural Heritage sem landsverkefni í hæsta forgangi og
skjalfest það með samstarfssamningi frá og með árinu 2005.
# text_en = It is no surprise that the federal and state governments have
classified the National Natural Heritage as a nation--wide endeavor of
the highest priority and have documented it starting in 2005
in a coalition contract.
```

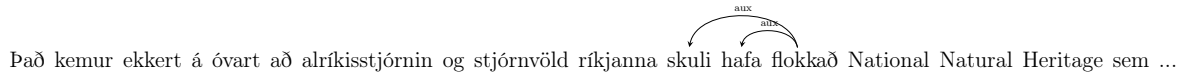


Figure A.20: Dependency relation: two auxiliaries in a verb predicate

A.2.2.4.2 cop

The copula definition has not been a part of the Icelandic grammar tradition. According to the UD guidelines, most European languages only have one word which can have the copula relation, i.e., 'to be'. Traditional grammar may label the verb

'to become' as copula but it is not according to the UD guidelines. The reason for treating the copula as the dependent of the predicate in the sentence, is that many languages are missing the explicit copula in the phrase–structure (Marneffe et al., 2014). There are six cases which must be considered when annotating copula. All of them are labeled with the copula relation except the existential meaning:

- Equation (aka identification): "hún er móðir mín"
- Attribution: "hún er ágæt"
- Location: "hún er inni í stofu"
- Possession: "hundurinn er hennar"
- Benefaction: "bókin er handa henni"
- Existence: "það er matur (í eldhúsinu)"

In Icelandic, the existential meaning often takes a compound particle, *að vera til*.

Figure A.21: Dependency relation: equation copula

```
# sent_id = n01003013
# text = Kannski voru kröfur um klæðaburð of strangar.
# text_en = Maybe the dress code was too stuffy.
```

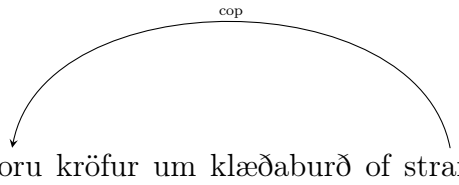


Figure A.22: Dependency relation: attribution copula

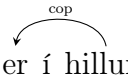


Figure A.23: Dependency relation: location copula

Figure A.24: Dependency relation: possession copula

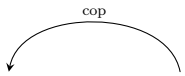


Figure A.25: Dependency relation: benefaction copula

A.2.2.4.3 mark

A marker is the word marking a clause as subordinate to another clause. It is also used for the infinitive marker *að*.

```
# sent_id = n04002020
# text = Það er erfitt að hugsa til þess að á Ítalíu séu færri kílómetrar af
neðanjarðarteinum en í Madrid.
# text_en = It is difficult to think that Italy has fewer km of underground line
than Madrid.
```

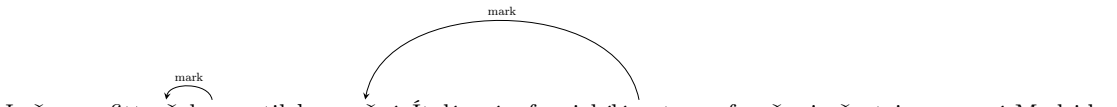


Figure A.26: Dependency relation: marker

A.2.3. Nominal Dependents

A.2.3.1. Nominals

A.2.3.1.1 nmod

nmod stands for nominal modifier and *nmod:poss* stands for a nominal modifier in possessive function.

```
# sent_id = n01061016
# text = En eftir því sem gagnrýnin á Clinton sem forsetafrú varð meinlegri jókst
hluttekning Karel.
```

```
# text_en = But, as the criticism of Clinton as First Lady became more caustic,
Karel became more empathetic.
```

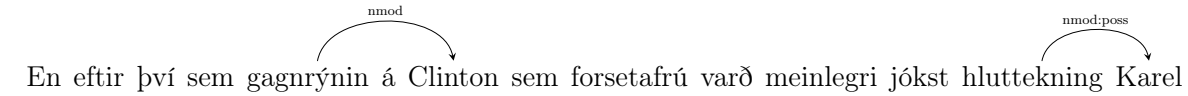


Figure A.27: Dependency relation: nominal and nominal possessive modifier

A.2.3.1.2 appos

Appositional modifier is used between two nominals to define, modify, name, or describe the first noun.

```
# sent_id = w02003037
# text = Árið 1832 seldi ríkið Württemberg verksmiðjueigandanum Georg Reichenbach
klaustrið fyrrverandi með því skilyrði að bómullarverksmiðja yrði sett á fót þar.
# text_en = In 1832, the former cloister was sold to the manufacturer
Georg Reichenbach by the state of Württemberg with the stipulation to
establish a cotton factory there.
```

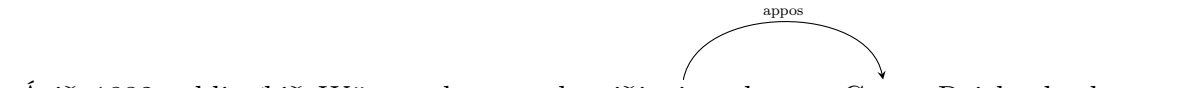


Figure A.28: Dependency relation: apposition

A.2.3.1.3 nummod

A numeric modifier of a noun is any number phrase that serves to modify the meaning of the noun with a quantity.

```
# sent_id = w02004021
# text = Uppruna sögulega ráðhússins í Obermarsberg má rekja til 13. aldar og
það var gert upp eftir þrjátíu ára stríðið.
# text_en = The historical city hall in Obermarsberg originates from the 13th
century and was refurbished after the Thirty Years' War.
```

Figure A.29: Dependency relation: numeric modifier

A.2.3.2. Clauses

A.2.3.2.1 acl

acl is an adjectival clause that modifies a nominal.

```
# sent_id = n01109008
# text = Kostnaðurinn breytist mánaðarlega og búist er við því að verðið hækki
á veturna, þegar notkun eykst.
# text_en = The cost will change monthly and the price is expected to rise in
winter as usage increases.
```

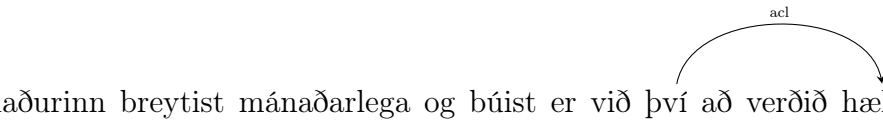


Figure A.30: Dependency relation: adjectival clause

A.2.3.3. Modifier Words

A.2.3.3.1 amod

amod is an adjectival modifier.

```
# sent_id = n01113021
# text = Eftir snilldarlega uppskurði og mikið af erfiðri endurhæfingu hef ´
ég náð mér að fullu.
# text_en = After some genius surgery and a lot of very tough rehab,
I have made a full recovery.
```

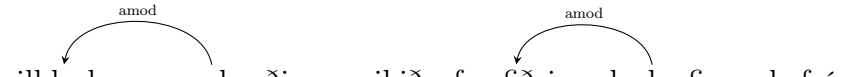


Figure A.31: Dependency relation: adjectival modifier

A.2.3.4. Function Words

A.2.3.4.1 det

det is a determiner relation between a nominal (head) and its determiner. In this corpus, all DET tags are labeled with the *det* dependency relation.

```
# sent_id = n01118010
# text = Samkvæmt talningu leikstjórans sjálfs hefur hann gert átta kvikmyndir
í fullri lengd til þessa.
# text_en = According to the director's own count, to date he has made eight
feature films.
```

Figure A.32: Dependency relation: determiner

A.2.3.4.2 case

The case relation is used for any case–marking element which is treated as a separate syntactic word like prepositions and postpositions. Case–marking elements are treated as dependents of the noun they attach to or introduce.

```
# sent_id = n01141002
# text = Á miðvikudaginn kynnti Microsoft nýtt spjallforrit fyrir vinnustaði,
sem er ætlað að keppa við toppforritið Slack.
# text_en = Microsoft announced on Wednesday a new workplace chat tool poised
to take on industry darling Slack.
```

Figure A.33: Dependency relation: case

A.2.4. Coordination

A.2.4.1. conj

conj is a conjunct relation between two elements connected by a coordinating conjunction. Conjunctions are treated asymmetrically: The head of the relation is the first conjunct and all the other conjuncts depend on it via the conj relation.

```
# sent_id = n01007012
# text = Hér er hægt að bera saman leiki og hversdagslíf okkar.
# text_en = There are parallels to draw here between games and our
everyday lives.
```

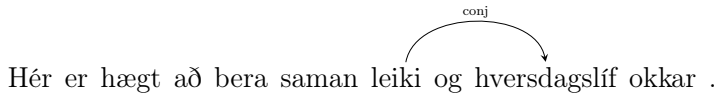


Figure A.34: Dependency relation: coordination

A.2.4.2. cc

A *cc* is the relation between a conjunct and a preceding coordinating conjunction.

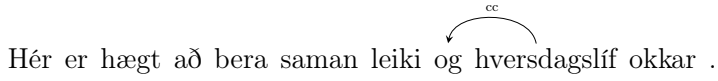


Figure A.35: Dependency relation: coordinating conjunction

A.2.5. Multiword Expressions (MWE)

A.2.5.1. fixed

The fixed relation is one of the three relations for multiword expressions (MWEs) (the other two being flat and compound). It is used for certain fixed grammaticized expressions that behave like function words or short adverbials. Fixed MWEs are annotated in a flat structure, where all subsequent words in the expression are attached to the first one using the fixed label. The assumption is that these expressions do not have any internal syntactic structure (except from a historical perspective) and that the structural annotation is in principle arbitrary. In Icelandic this mostly

applies to subordinating conjunctions in adverbial and relative clauses: *svo að, af því að, því að, þar eð, þar sem, fyrir því að, með því að, úr því að, sökum þess að, sakir þess að, vegna þess að, áður en, á meðan, eftir að, frá því að, frá því er, fyrr en, jafnskjótt og, jafnskjótt sem, óðar en, undireins og, strax og, um leið og, þangað til að, þar til er, þá er, til þess er, þar til að, svo framarlega sem, þó að, enda þótt, jafnvel þótt, þrátt fyrir (það) að, til þess að, til að, eins og, svo sem, heldur en, þeim mun, þar sem, þar er, þangað sem, þangað er, þaðan sem, þaðan er, hvert sem, hvert er, hvar sem, hvaðan sem, hvaðan er, hvernig sem, hvenær sem*

```
# sent_id = n01014012
# text = Fyrirheitið um nýjar pantanir Konunglega sjóhersins til að tryggja
skipasmíðaiðnaðinn í Clyde var gefið fyrir þjóðaratkvæðagreiðsluna um sjálfstæði
Skotlands árið 2014.
# text_en = The promise of new Royal Navy orders to secure the Clyde shipbuilding
industry was made before the Scottish independence referendum in 2014.
```

Fyrirheitið um nýjar pantanir Konunglega sjóhersins til (fixed) að tryggja skipasmíðaiðnaðinn ...

Figure A.36: Dependency relation: fixed relation I

```
# sent_id = n01016032
# text = Mate 9 símana skortir gervigreindarviðmót, svo sem Google Assistant eða
Siri frá Apple.
# text_en = The Mate 9 phones lack an artificial intelligence interface, like
the Google Assistant or Apple's Siri.
```

Mate 9 símana skortir gervigreindarviðmót , svo (fixed) sem Google Assistant eða Siri frá Apple .

Figure A.37: Dependency relation: fixed relation II

```
# sent_id = n01018024
# text = Þetta er stundum eins og ofurkraftur.
# text_en = It's like a super power sometimes.
```

Þetta er stundum eins (fixed) og ofurkraftur .

Figure A.38: Dependency relation: fixed relation III

```
# sent_id = w01003056
# text = Á meðal graníteyja eru Seychelleyjar og Tioman--eyja og eldfjallaeyjar
á borð við Sankti Helenu.
# text_en = Granite islands include Seychelles and Tioman and volcanic islands
such as Saint Helena.
```

Figure A.39: Dependency relation: fixed relation IV

A.2.5.2. flat

flat is used for exocentric (headless) semi–fixed MWEs like names (Hildur Jónsdóttir) and dates (1. janúar 2020). It contrasts with fixed, which applies to completely fixed grammaticized MWEs (like *þrátt fyrir að*, and with compound, which applies to endocentric (headed) MWEs (like *forseti Alþingis*).

```
# sent_id = w01068027
# text = 9. september 2002 var Aldrin lokkaður á hótel í Beverly Hills undir því
yfirskini að taka viðtal við hann um geimferðir fyrir japanskan barnaþátt.
# text_en = On September 9, 2002, Aldrin was lured to a Beverly Hills hotel on
the pretext of being interviewed for a Japanese children's television show
on the subject of space.
```

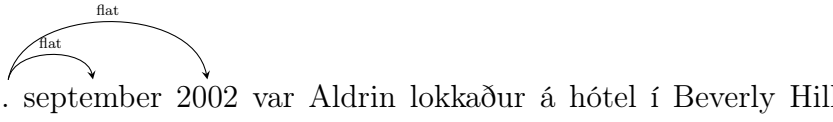


Figure A.40: Dependency relation: flat relation

A.2.5.3. flat:name

flat:name is a relation for flat structure as described above and applies specifically to proper names.

```
# sent_id = w01073054
# text = Ritstjórinn Stephen Wheeler hélt Kipling við efnið, en skrifþörf
Kiplings var óstöðvandi.
# text_en = Stephen Wheeler, the editor, worked Kipling hard, but Kipling's
need to write was unstoppable.
```

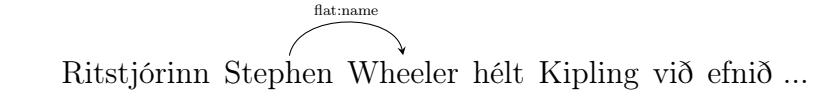


Figure A.41: Dependency relation: flat:name relation

A.2.5.4. flat:foreign

flat:foreign is used to label sequences of foreign words. These are given a linear analysis: the head is the first token in the foreign phrase. *flat:foreign* does not apply to loanwords or to foreign names. It applies to quoted foreign text incorporated in a sentence/discourse of the host language. This structure occurs in IcePaHC but not in the Icelandic PUD.

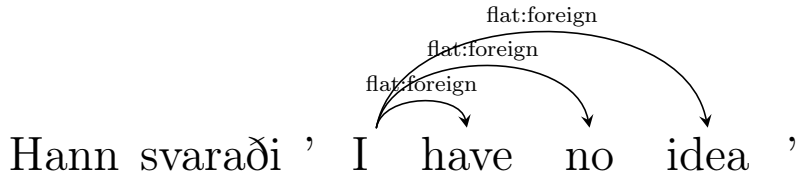


Figure A.42: Dependency relation: flat:foreign relation

A.2.5.5. compound

compound is used for any kind of compounding: noun compounds (e.g., phone book), but also verb and adjective compounds that are more common in other languages. This construction is not frequent in the Icelandic PUD and resembles the flat:name structure however the *compound* is especially applied to institutions, companies, brands etc. that have assembled names. When in doubt, the annotation done for the Swedish and English PUD was followed.

```
# sent_id = n01016032
# text = Mate 9 símana skortir gervigreindarviðmót, svo sem Google Assistant eða
Siri frá Apple.
# text_en = The Mate 9 phones lack an artificial intelligence interface, like the
Google Assistant or Apple’s Siri.
```

Figure A.43: Dependency relation: compound relation

A.2.5.6. compound:prt

The *compound:prt* is a separable verb particle and is used to mark the separated particles of particle verbs. It is a subtype of the compound relation.

```
# sent_id = n01020021
# text = Og nú þegar stefnir í að Kína verði stærsti flugmarkaður heims á næsta
áratug gefur sýningin Beijing færi á að sýna metnað sinn í almenningsflugi,
auk varnarmála.
# text_en = And with China set to become the world's biggest aviation market
in the next decade, the show is an opportunity for Beijing to demonstrate
its ambitions in civil aviation as well as defence.
```

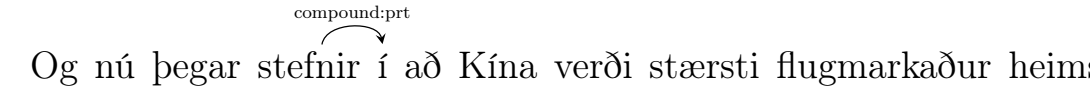


Figure A.44: Dependency relation: compound:prt relation I

```
# sent_id = n01030008
# text = Samkvæmt lögregludeild borgarinnar brutust út slagsmál þegar
„stór hópur fólks ... sem sagðist vera valdhafar á svæðinu“ kom á
tökustað á þriðjudaginn.
# text_en = According to the city's police department, a fight broke out after
"a large group of people... who identified themselves as district
authorities" arrived on set on Tuesday.
```

Figure A.45: Dependency relation: compound:prt relation II

A.2.6. Loose

A.2.6.1. parataxis

Loose joining relations are identified with two labels, *list* and *parataxis*. List annotation is used in the Swedish and Danish UD treebanks but not the Norwegian nor the Faroese versions. For now, the list labelling is not applied in the Icelandic annotation. The *parataxis* relation is a relation between a word and other elements, such as a sentential parenthetical or a clause after a ":" or a ";", placed side by side without any explicit coordination, subordination, or argument relation with the head word.

```
# sent_id = n01035025
# text = „Við teljum ekki að hinn grunaði hafi tengsl við þennan skóla,
stúlkurnar tvær eða Abbotsford-svæðið sérstaklega“ sagði hún.
# text_en = "We do not believe the suspect has ties to this school, or to the
two girls, or specifically to the Abbotsford area," she said.
```

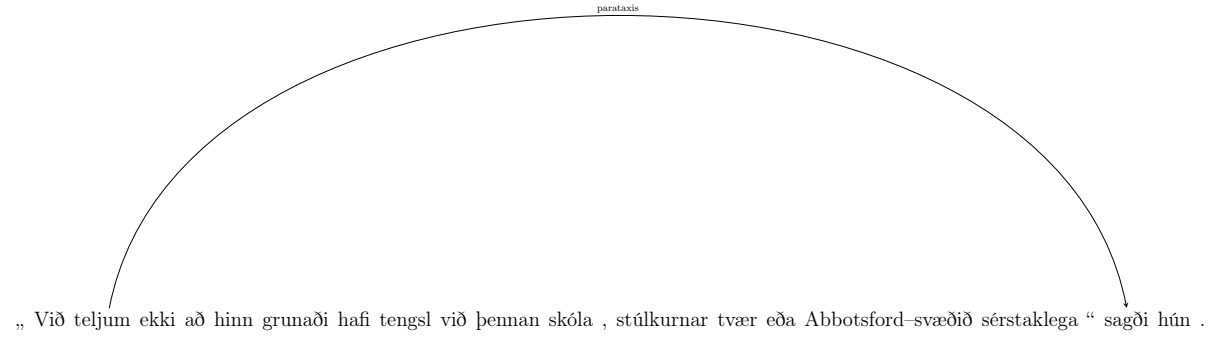


Figure A.46: Dependency relation: parataxis relation

A.2.7. Other

A.2.7.1. punct

punct stands for any punctuation in a clause.

```
# sent_id = n01035025
# text = ,,Við teljum ekki að hinn grunaði hafi tengsl við þennan skóla, stúlkurnar
tvær eða Abbotsford--svæðið sérstaklega‘‘ sagði hún.
# text_en = "We do not believe the suspect has ties to this school, or to the two
girls, or specifically to the Abbotsford area," she said.
```

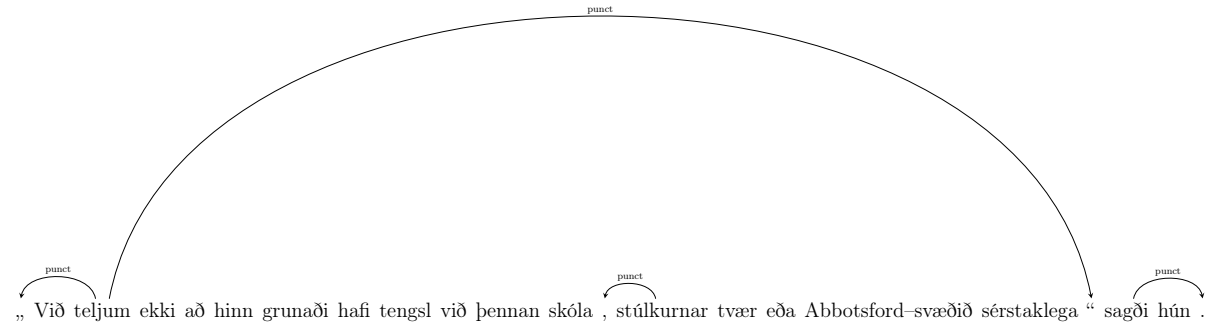


Figure A.47: Dependency relation: punctuation relation

A.2.7.2. root

The *root* relation points to the root of the sentence and is a fake grammatical node. It is indexed with 0 since the indexing of real words in the sentence starts at 1. There can only be one root in every tree.

```
# sent_id = n01044009
# text = Það leiddi til uppljóstrana á síðustu tveimur dögum um að minnst sex
aðrir blaðamenn í Quebec hafi verið undir eftirliti héraðslögreglunnar.
# text_en = That led to revelations over the last two days that at least
six other Quebec journalists were targeted by provincial police
surveillance operations.
```

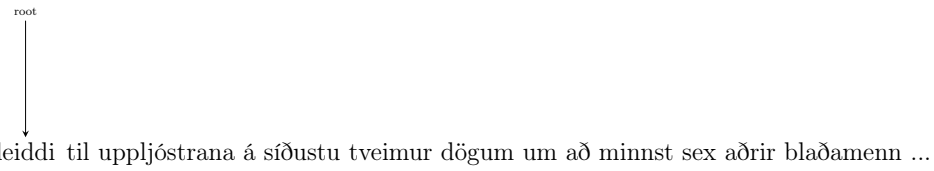


Figure A.48: Dependency relation: root relation

A.2.7.3. dep

The *dep* relation is used for unspecified dependency. The use of *dep* should be avoided but can occur when in strange grammatical constructions or caused by limitations in conversion or parsing software. There are no *dep* relations in the Icelandic PUD.

B. Mapping of IFD tagset to CoNLL-U format

As explained in section 3.3, both the UPOS and the FEATS fields in the CoNLL-U format are extracted from the IFD tagset. An example of one Icelandic sentence that has been tagged and lemmatized is given in table B.1 and then the same sentence is displayed in B.2 after being mapped from IFD tagset to UPOS and FEATS and converted to CoNLL-U format (first 6 fields); ID, TOKEN, LEMMA, UPOS, XPOS and FEATS. The details of the mapping process are listed below.

Token	IFD-tag	Lemma
Greining	nven	greining
leiddi	sfg3eþ	leiða
í	ao	í
ljós	nheo	ljós
að	c	að
sögulegt	lhensf	sögulegur
magn	nhen	magn
gagna	nhfe	gagn
hafði	sfg3eþ	hafa
verið	ssg	vera
notað	sþghen	nota
til	ae	til
að	cn	að
valda	sng	valda
trufluninni	nveþg	truflun
.	.	.

Table B.1: Tagged and Lemmatized Icelandic sentence

ID	FORM	LEMMA	UPOS	XPOS	FEATS
1	Greining	greining	NOUN	nven	Case=Nom|Definite=Ind|Gender=Fem|Number=Sing
2	leiddi	leiða	VERB	sfg3eþ	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act
3	í	í	ADP	ao	
4	ljós	ljós	NOUN	nheo	C̄ase=Acc|Definite=Ind|Gender=Neut|Number=Sing
5	að	að	SCONJ	c	
6	sögulegt	sögulegur	ADJ	lhensf	C̄ase=Nom|Definite=Ind|Degree=Pos|Gender=Neut|Number=Sing
7	magn	magn	NOUN	nhen	Case=Nom|Definite=Ind|Gender=Neut|Number=Sing
8	gagna	gagn	NOUN	nhfe	Case=Gen|Definite=Ind|Gender=Neut|Number=Plur
9	hafði	hafa	VERB	sfg3eþ	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act
10	verið	vera	AUX	ssg	VerbForm=Sup|Voice=Act
11	notað	nota	VERB	sþghen	Case=Nom|Gender=Neut|Number=Sing|Tense=Past|VerbForm=Part|Voice=Act
12	til	til	ADP	ae	_
13	að	að	PART	cn	
14	valda	valda	VERB	sng	V̄erbForm=Inf|Voice=Act
15	trufluninni	truflun	NOUN	nveþg	Case=Dat|Definite=Def|Gender=Fem|Number=Sing
16	.	.	PUNCT	.	_

Table B.2: Icelandic sentence in CoNLL-U format (fields 1-6)

IFD column	IFD symbol	UD UPOS	UD Features
1	n	NOUN	
2	k		Gender=Masc
2	v		Gender=Fem
2	h		Gender=Neut
3	e		Number=Sing
3	f		Number=Plur
4	n		Case=Nom
4	o		Case=Acc
4	þ		Case=Dat
4	e		Case=Gen
5	g		Definite=Det
5	<empty>		Definite=Ind
6	m, ö, s	PROPN	
1	s	VERB	
2	f		Mood=Ind, VerbForm=Fin
2	n		VerbForm=Inf
2	b		Mood=Imp, VerbForm=Fin
2	v		Mood=Sub, VerbForm=Fin
2	s		VerbForm=Sup
2	l		VerbForm=Part, Tense=Pres
2	þ		(see next section for past participle)
3	g		Voice=Act
3	m		Voice=Mid
4	1		Person=1
4	2		Person=2
4	3		Person=3
5	e		Number=Sing
5	f		Number=Plur
6	n		Tense=Pres
6	þ		Tense=Past
1	s	VERB	(past participle)
2	þ		VerbForm=Part, Tense=Past
3	g		Voice=Act
3	m		Voice=Mid
4	k		Gender=Masc
4	v		Gender=Fem
4	h		Gender=Neut
5	e		Number=Sing
5	f		Number=Plur
6	n		Case=Nom
6	o		Case=Acc

IFD column	IFD symbol	UD UPOS	UD Features
1	l	ADJ	
2	k		Gender=Masc
2	v		Gender=Fem
2	h		Gender=Neut
3	e		Number=Sing
3	f		Number=Plur
4	n		Case=Nom
4	o		Case=Acc
4	þ		Case=Dat
4	e		Case=Gen
5	s		Definite=Det
5	v		Definite=Ind
6	f		Degree=Pos
6	m		Degree=Cmp
6	e		Degree=Sup
1	a		
2	a	ADV	
2	u	INTJ	
2	o,þ,e	ADP	
2	o,þ,e	ADP	
2	m,e		ADV
1	f	PRON	
2	a		PronType=Dem
2	b		PronType=Ind
2	e		Poss=Yes
2	o		PronType=Ind
2	p		PronType=Prs
2	s		PronType=Int
2	t	SCONJ	
3	k		Gender=Masc
3	v		Gender=Fem
3	h		Gender=Neut
3	1		Person=1
3	2		Person=2
4	e		Number=Sing
4	f		Number=Plur
6	n		Case=Nom
6	o		Case=Acc
6	þ		Case=Dat
6	e		Case=Gen

IFD column	IFD symbol	UD UPOS	UD Features
1	c	CCONJ	(if token matches: ’og’, ’eða’, ’en’, ’heldur’, ’enda’, ’ellegar’, ’bæði’, ’hvorki’, ’né’, ’annaðhvort’, ’eða’, ’ýmist’)
1	c	SCONJ	
2	p	SYM	
2	n	PART	
2	t	SCONJ	
1	g	DET	
2	k		Gender=Masc
2	v		Gender=Fem
2	h		Gender=Neut
3	e		Number=Sing
3	f		Number=Plur
4	n		Case=Nom
4	o		Case=Acc
4	þ		Case=Dat
4	e		Case=Gen
1	t	NUM	
2	f		NumType=Card
3	k		Gender=Masc
3	v		Gender=Fem
3	h		Gender=Neut
4	e		Number=Sing
4	f		Number=Plur
5	n		Case=Nom
5	o		Case=Acc
5	þ		Case=Dat
5	e		Case=Gen
1		PUNCT	(token matches:’„’, PunctSide=Ini, token matches:”“’, PunctSide=Fin)
1		PUNCT	(if token matches: ’.’, ’,’, ’:’, ’;’, ’(’, ’)’, ’!’, ’?’, ’...’, ’[’, ’]’, ’‟’)
1	e	PROPN	
1	x	X	

Bibliography

Ahrenberg, L. (2015). Converting an English-Swedish Parallel Treebank to Universal Dependencies. In Nivre, J. and Hajičová, E., 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 Hartmann, M. and Plank, B., editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics NODALIDA-2019*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Barry, J., Wagner, J., and Foster, J. (2019). Cross–lingual Parsing with Polyglot Training and Multi-treebank Learning: A Faroese Case Study. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*, pages 163—-174. Association for Computational Linguistics.

Berdicevskis, A., Çöltekin, Ç., Ehret, K., von Prince, K., Ross, D., Thompson, B., Yan, C., Demberg, V., Lupyan, G., Rama, T., and Bentz, C. (2018). Using universal dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17, Brussels, Belgium. Association for Computational Linguistics.

Bick, E. (2003). Arboretum, a Hybrid Treebank for Danish. In *Proceedings of the 2ndWorkshop on Treebanks and Linguistic Theories*.

Bilgram, T. and Keson, B. (1998). The Construction of a Tagged Danish Corpus. In *Proceedings of the 11th Nordic Conference of Computational Linguistics (NODALIDA 1998)*, pages 129–139. Center for Sprogteknologi, University of Copenhagen.

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.

Bjarnadóttir, K. (2016). Setningafræði orðabókarmanns: ÍSLEX-venslamálfræði [The Syntax of a Lexicographer: ISLEX Dependency Grammar]. presentation at a conference in honor of Höskuldur Þráinsson's 70th birthday, Árnagarður, Reykjavík, January 16, 2016. pages 441–483.

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

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The tiger treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Candito, M., Evang, K., Oepen, S., and Seddah, D., editors (2019). *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, Paris, France. Association for Computational Linguistics.

Eisner, J. M. (1996). Three New Probabilistic Models for Dependency Parsing: An Exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Faarlund, J. T., Lie, S., and Vannebo, K. I. (1997). *Norsk referansegrammatikk*. Universitetsforlaget.

Ganchev, K. and Das, D. (2013). Cross-Lingual Discriminative Learning of Sequence Models with Posterior Regularization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1996–2006. Association for Computational Linguistics.

Groß, T. (2014). Clitics in dependency morphology. In Gerdes, K., Hajicova, E., and Wanner, L., editors, *Dependency Linguistics: Recent advances in linguistic theory using dependency structures*, pages 229–252, Amsterdam, Netherlands. John Benjamins Publishing Company.

Helgadóttir, S. (2005). Testing Data–Driven Learning Algorithms for PoS Tagging of Icelandic. In Holmboe, H., editor, *Nordisk Sprogteknologi 2004*. Museum Tusculanums Forlag.

Ingason, A., Loftsson, H., Rögnvaldsson, H., Sigurðsson, E. F., and Wallenberg, J. (2014). Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. pages 789–795.

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. (2019). Nefnir: A high accuracy lemmatizer for icelandic. In Hartmann, M. and Plank, B., editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics NODALIDA-2019*, pages 310–315. Linköping University Electronic Press.

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

Johansson, R. and Nugues, P. (2007). Extended Constituent-to-dependency Conversion for English. In *Proceedings of NODALIDA 2007*, pages 105–112, Tartu, Estonia.

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

Kakkonen, T. (2006). Dependency Treebanks: Methods, Annotation Schemes and Tools. *CoRR*, abs/cs/0610124.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Keith Percival, W. (1990). Reflections on the History of Dependency Notions in Linguistics. *Historiographia Linguistica*, 17:29–47.

Kiss, T. and Alexiadou, A. (2015). *Syntax - Theory and Analysis*. Walter de Gruyter GmbH & Co KG.

Kromann, M. T. and Lynge, S. K. (2004). *Danish Dependency Treebank v. 1.0*. Department of Computational Linguistics, Copenhagen Business School.

Kárason, Ö. (2016). Athugun á fýsileika sjálfvirkrar vörpunar liðgerðargreiningu íslensks trjábanka yfir í venslagreiningu. Unpublished.

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency parsing*. Number 2 in Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

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

Loftsson, H. and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite–State Parser for Icelandic. In Nivre, J., Kaalep, H.-J., Muischnek, K., and Koit, M., editors, *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, pages 128–135.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592. European Languages Resources Association (ELRA).

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non–Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Mrini, K., Dernoncourt, F., Bui, T., Chang, W., and Nakashole, N. (2019). Rethinking Self–Attention: An Interpretable Self–Attentive Encoder–Decoder Parser.

Nikulásdóttir, A., Guðnason, J., and Steingrímsson, S. (2017). Language Technology for Icelandic 2018-2022: Project Plan. Technical report, Mennta- og menningarmálaráðuneytið, Reykjavík.

Nivre, J. (2003). An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)*, pages 149–160.

Nivre, J. (2005a). Dependency Grammar and Dependency Parsing. Technical Report MSI 05133, Växjö University, School of Mathematics and Systems Engineering.

Nivre, J. (2005b). *Inductive Dependency Parsing of Natural Language Text*. Phd, Växjö University.

Nivre, J., Abrams, M., Agić, Ž., and et al. (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.

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., 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 Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., 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., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*.

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

Øvrelid, L. and Hohle, P. (2016). Universal Dependencies for Norwegian. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., 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 Savary, A., Ramisch, C., Hwang, J. D., Schneider, N., Andresen, M., Pradhan, S., and Petruck, M. R. L., 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.

Petrov, S., Das, D., and McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Pind, J., Magnússon, F., and Briem, S. (1991). *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland.

Polguère, A. and Melcuk, I. (2009). *Dependency in Linguistic Description*. Studies in Language Companion Series. John Benjamins Publishing Company.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., and Wallenberg, J. (2012a). The Icelandic Parsed Historical Corpus (IcePaHC). In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., 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*. META-NET.

Rúnarsson, K. (2017). Samba: Automatic Identification of Verbal Expressions in Icelandic. Master's thesis, University of Iceland.

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannessen, J. (2014). The Norwegian Dependency Treebank. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S.,

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., Kárason, Ö., and Loftsson, H. (2019). Augmenting a BiLSTM tagger with a Morphological Lexicon and a Lexical Category Identification Step. In Mitkov, R., Angelova, G., and Bontcheva, K., editors, *Proceedings of Recent Advances in Natural Language Processing*, pages 1162–1169, Varna, Bulgaria.

Straka, M., Hajič, J., Strakova, J., and Hajič jr, J. (2015). Parsing Universal Dependency Treebanks using Neural Networks and Search–Based Oracle. In Dickinson, M., Hinrichs, E., Patejuk, A., and Przepiórkowski, A., editors, *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 208–220.

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

Tesniere, L. (2015). *Elements of Structural Syntax*. John Benjamins.

Tiedemann, J., Agić, Ž., and Nivre, J. (2014). Treebank Translation for Cross-Lingual Parser Induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140. Association for Computational Linguistics.

Tyers, F. M., Sheyanova, M., Martynova, A., Stepachev, P., and Vinogradovsky, K. (2018). Multi–source synthetic treebank creation for improved cross-lingual dependency parsing. In de Marneffe, M.-C., Lynn, T., and Schuster, S., 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. (2017). UD annotatrix: An annotation tool for Universal Dependencies. In Hajič, J., editor, *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 10–17, Prague, Czech Republic.

Volk, M., Marek, T., and Samuelsson, Y. (2018). *Annotation, exploitation and evaluation of parallel corpora: TC3 I*. Language Science Press, Berlin.

Wang, X., Pham, H., Yin, P., and Neubig, G. (2018). A tree–based Decoder for Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4772–4777, Brussels, Belgium. Association for Computational Linguistics.

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.

Yurafsky, D. and Martin, J. H. (2018). Speech and Language Processing. Unpublished, 3rd ed.draft.

Zeman, D. (2017). Core Arguments in Universal Dependencies. In Montemagni, S. and Nivre, J., 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.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

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*.

Øvrelid, L., Kåsen, A., Hagen, K., Nøklestad, A., Solberg, P. E., and Johannessen, J. B. (2018). The LIA Treebank of Spoken Norwegian Dialects. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., 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 Mitkov, R., Angelova, G., and Bontcheva, K., editors, *Proceedings of Recent Advances in Natural Language Processing*, pages 1397–1404, Varna, Bulgaria.

Þráinsson, H. (2007). *The Syntax of Icelandic*. Cambridge Syntax Guides. Cambridge University Press.