REYKJAVÍK UNIVERSITY
HÁSKÓLINN Í REYKJAVÍK

A shallow syntactic annotation scheme for Icelandic text

Hrafn Loftsson, Eiríkur Rögnvaldsson

RUTR-SSE06004, December 2006
Department of Computer Science

Reykjavík University - School of Science and Engineering

Technical Report

ISSN 1670-5777

A shallow syntactic annotation scheme for Icelandic text

Hrafn Loftsson*, Eiríkur Rögnvaldsson†

School of Science and Engineering – Department of Computer Science
Technical Report RUTR-SSE06004 – December 2006

Abstract: We describe a shallow syntactic annotation scheme for Icelandic text. The scheme comprises a set of grammatical descriptors and their application guidelines. The descriptors consist of brackets and labels which indicate constituent structure and functional relations. Additionally, we describe a grammar definition corpus, annotated using the annotation scheme. The annotation scheme has been developed as a part of a shallow parsing project.

Keywords: Annotation scheme; Grammar definition corpus; Constituent structure; Syntactic functions; Shallow parsing.

(Útdráttur: næsta síða)

This report is part of the project *Shallow parsing of Icelandic text*, which was partially supported by Rannís (The Icelandic Centre for Research), grant 060010021.

* Reykjavík University, Ofanleiti 2, IS-103 Reykjavík, Iceland. hrafn@ru.is
† University of Iceland, Suðurgötu, IS-101, Reykjavík, Iceland. eirikur@hi.is

HÁSKÓLINN Í REYKJAVÍK
REYKJAVÍK UNIVERSITY

Grunnt þáttunarskema fyrir íslenskan texta

Hrafn Loftsson, Eiríkur Rögnvaldsson

Tækni- og verkfræðideild – Tölvunarfræðisvið
Tækniskýrsla RUTR-SSE06004 – Desember 2006

Útdráttur: Við lýsum grunnu þáttunarskema fyrir íslenskan texta. Skemað samanstendur af mengi af málfræðilýsingum ásamt leiðbeiningum um notkun þeirra. Lýsingarnir felast í hornklofum og mörkum sem sýna formgerð setninga og setningafræðileg hlutverk. Við lýsum jafnframt skilgreiningarmálheild sem hefur verið mörkuð með því að nota þáttunarskemað. Þáttunarskemað var þróað sem hluti af verkefni á sviði hlutaþáttunar.

Lykilorð: Þáttunarskema; Skilgreiningarmálheild; Setningaliðir; Setningafræðileg hlutverk; Hlutaþáttun.

(Abstract: previous page)

Contents

1 Introduction

In this paper, we describe a shallow syntactic annotation scheme for Icelandic text. The annotation scheme has been developed as a part of a shallow parsing project.

After reviewing the EAGLES guidelines for syntactic annotation and different types of annotation schemes used in treebanks, we propose an annotation scheme for Icelandic text. Our scheme comprises a set of grammatical descriptors and their application guidelines. The grammatical descriptors consist of brackets and labels which indicate constituent structure and functional relations. When describing the grammatical descriptors we show examples (general principles) of their use. Additionally, we have constructed a *grammar definition corpus*, a set of carefully chosen sentences annotated using the annotation scheme.

Our annotation scheme is specifically designed for the purpose of being used in shallow parsers, and it is the first such scheme published for the Icelandic language[1].

2 Related work

2.1 The EAGLES guidelines

In 1996, EAGLES proposed guidelines for syntactic annotation of corpora (EAGLES, Expert Advisory Group for Language Engineering Standards 1996). In the report, syntactic annotation is defined as *the practise of adding syntactic information to a corpus by incorporating into the text indicators of syntactic structure: e.g. labeled bracketing, or symbols indicating dependency relations between words*[2].

In the EAGLES proposals, the following layers of information are recognised (which may or may not be encoded in a particular syntactic annotation scheme):

1. **Bracketing of segments**. Involves the delimitation of segments, usually with square brackets, which are recognised as having a syntactic integrity (sentences, clauses, phrases, words).

[1] Our annotation scheme is based on traditional syntactic analysis for Icelandic as, for example, presented in (Þráinsson 1999).

[2] In our discussion, we do not distinguish between adding syntactic information to a corpus vs. parsing running text.

2. **Labeling of segments**. Indication of formal category of the constituents identified by the bracketing, such as a noun phrase, a verb phrase, a prepositional phrase, etc.

3. **Showing dependency relations**. Head-dependent relations between words, e.g. adjectives and the nouns they modify. Usually shown with dependency trees: a set of arrows pointing from a head to a dependent (or *vice verse*).

4. **Indicating functional relations**. Labeling of segments according to their syntactic function, such as subject, object, predicate, etc.

5. **Marking subclassification of syntactic segments**. Assigning feature values to phrases or words, e.g. marking a noun phrase as singular or a verb phrase as past tense.

6. **Deep or logical information**. This includes a variety of syntactic phenomena, such as co-referentiality, cross-reference and syntactic discontinuity.

7. **Information about the rank of a syntactic unit**. This is obtainable from most parser outputs by the embedding of marked brackets.

8. **Special syntactic characteristics of spoken language**. Indication of false starts, reiterations, pauses etc.

For the purpose of shallow syntactic annotation, the main emphasis is on bracketing and labeling of segments and indication of functional relations (i.e. items 1,2 and 4 above). In some cases, shallow dependency relations are shown as well. Subclassification information of syntactic segments is usually not marked by shallow parsers. The reason is probably that, in many cases, the part-of-speech (POS) tags contain morphological information which can be used to derive subclassification information, like the gender and number of a noun phrase.

The emphasis on partial information entails that a shallow annotation scheme does not include "deep" information. To date, main emphasis has been put on parsing written text and, thus, annotations generated by parsers, generally, do not include indication of special spoken language characteristics.

2.2 Treebanks

A treebank is a syntactically annotated corpus, in which the annotations follow a particular annotation scheme. Most treebanks have been built by manually, or semi-automatically, adding syntactic annotations to a POS tagged corpus. Treebanks

have, for example, been used to facilitate linguistic research, as training corpora for data-driven methodologies and as evaluation resources for parsers. Three main kinds of annotation are used in practise: annotation of constituent structure, annotation of functional structure (syntactic/grammatical functions) and theory-specific annotation (Nivre 2002). The annotation found in most treebanks is, in fact, a combination of two or even three of theses categories.

The annotation of constituent structure (or bracketing) is the most common annotation method. It is, for example, used in the well known Penn Treebank (Marcus et al. 1993). Usually, this kind of annotation consists of POS tags for individual words, augmented with annotation of major constituents, like noun phrases, prepositional phrases, verb phrases, etc. These schemes are *usually intended to be theory-neutral and therefore try to use mostly uncontroversial categories that are recognised in all or most syntactic theories that assume some notion of constituent structure* (Nivre 2002). Thus, the advantage with this annotation method is that the treebank can be used by a larger group of researchers working within different theoretical frameworks. The disadvantage, however, is the risk that the annotation contains too little information, which makes the treebank inadequate to use for anyone.

In recent years, annotation of functional structure has become increasingly important. First, grammatical function annotation has been added to many corpora annotated with constituent structure, e.g. the Penn Treebank II (Marcus et al. 1994). Secondly, so called dependency syntax annotation schemes (signifying dependencies between words) have been developed, in which dependency structure is added directly on top of morphological information without any bracketing. The Prague Dependency Treebank of Czech is probably the best known example of this type of annotation structure (Hajič 1998).

The third kind of annotation scheme is the one which uses representations from a particular grammatical theory, for example HPSG (Pollard and Sag 1994), to annotate sentences. The advantage with theory-supporting treebanks is that they are more useful for people working with the selected type of grammatical theory, but the disadvantage is that they are not as appropriate for people that do not use the specific theoretical framework.

2.3 A grammar definition corpus

When designing an annotation scheme, it can be helpful to create a *grammar definition corpus* (GDC), *a representative collection of utterances consistently analysed using a fixed set of grammatical descriptors* (Voutilainen 1997). Ideally, such a corpus

should provide unambiguous answers to questions on how to annotate any sentence in the given language.

The GDC can, moreover, be used in the development phase of a parser, since, the parser should, preferably, be able to produce (almost) equivalent annotation when annotating sentences in this corpus.

3 An annotation scheme for Icelandic text

In this section, we propose a shallow annotation scheme for Icelandic text. By shallow annotation, we mean that syntactic structures are rather flat and simple, i.e. the main emphasis is to annotate core phrases without showing a complete parse tree.

With reference to the EAGLES guidelines, our scheme consists of brackets and labels indicating constituent structure and functional relations (syntactic functions). Our scheme, thus, follows the dominant paradigm in treebank annotation, i.e. it *is the kind of theory-neutral annotation of constituent structure with added functional tags* (Nivre 2002).

Our reason for developing a shallow annotation scheme (as opposed to a full/deep annotation scheme) is that the scheme is being used in a shallow parsing project. Our parser is an incremental finite-state parser (Grefenstette 1996), a sequence of transducers each of which adds syntactic information, such as brackets and names for grammatical functions, into the text. Finite-state parsers are effective because they are just a pipeline of lexical analysers, and their aim is not to consider all possible analysis of a given sentence, as is the case for full parsers. Furthermore, these parsers are robust because they are not as sensitive to (grammatical) errors in the text as parsers based on full parsing methods. The reason is that full parsers sometimes reject correct analysis of a sentence part on lower levels in the parse tree on the ground that it does not fit into a global parse (Abney 1996). Shallow parsing techniques do not have these problems because their aim is *to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis* (Abney 1996).

We assume that the text to be annotated has already been POS tagged using the tagset created in the making of the Icelandic Frequency Dictionary (IFD) corpus (Pind et al. 1991). This tagset includes both word class and morphological information.

At the end of this section, we describe a GDC annotated using our scheme.

3.1 Constituent structure

The EAGLES guidelines recommends annotation of the following constituent categories: sentence, clause, noun phrase, verb phrase, adjective phrase, adverb phrase and prepositional phrase. Since our annotation scheme puts emphasis on core phrases, we neither include sentence nor clause categories.

We use brackets and labels to indicate constituents. Two labels are attached to each marked constituent: the first one denotes the beginning of the constituent, the second one denotes the end (e.g. [NP ...NP]).

The main labels are **AdvP**, **AP**, **NP**, **PP** and **VP** – the standard labels used for syntactic annotation (denoting adverb, adjective, noun, prepositional and verb phrase, respectively). Additionally, we use the labels **CP**, **SCP**, **InjP**, **APs**, **NPs** and **MWE** for marking coordinating conjunctions, subordinating conjunctions, interjections, a sequence of adjective phrases, a sequence of noun phrases, and multiword expressions, respectively. Hence, in our scheme, every word is a part of some constituent structure.

In the following sections, we describe the structure of each constituent in more detail. For each constituent, we show examples obtained from our GDC. The English gloss (most often a word-by-word translation) appears in parenthesis with most of the examples. For saving space, we leave out the POS tag associated with each word in the examples.

3.1.1 Multiword expression phrases

A multiword expression (MWE) phrase comprises fixed multiword expressions which function as a single word. We distinguish between four kinds of MWEs, i.e. expressions that function as i) a conjunction (MWE_CP), ii) an adverb (MWE_AdvP), iii) an adjective (MWE_AP), and iv) a preposition (MWE_PP).

Below we show 2-3 examples of each kind:

1. [MWE_CP eins og MWE_CP] (as)
2. [MWE_CP til að MWE_CP] (in order to)
3. [MWE_CP á meðan MWE_CP] (while)
4. [MWE_AdvP hvers vegna MWE_AdvP] (why)
5. [MWE_AdvP allt í einu MWE_AdvP] (suddenly)
6. [MWE_AdvP til dæmis MWE_AdvP] (for example)

7. [MWE_AP alls konar MWE_AP] (all kinds of)
8. [MWE_AP hvers kyns MWE_AP] (every kind of)
9. [MWE_PP fyrir framan MWE_PP] (in front of)
10. [MWE_PP út í MWE_PP] (out into)
11. [MWE_PP innan um MWE_PP] (among)

In example no. 9, the preposition ("fyrir") precedes the adverb ("framan"), but in examples no. 10 and 11 the adverbs ("út", "innan") precede the prepositions ("í", "um").

We have compiled a list of multiword expressions for each of the different kinds of MWEs (see Appendix A).

3.1.2 Adverb phrases

An adverb phrase ([AdvP ... AdvP]) consists of a sequence of one or more adverbs. The following are examples of adverb phrases:

1. [AdvP ekki AdvP] (not)
2. [AdvP svo AdvP] (so)
3. [AdvP þar AdvP] (there)
4. [AdvP þó AdvP] (although)
5. [AdvP þar með AdvP] (thereupon)
6. [AdvP í gær AdvP] (yesterday)
7. [AdvP þá fyrst AdvP] (then first)
8. [AdvP ekki síst AdvP] (not least)

Note that two (or more) adjacent adverbs are not necessarily part of the same adverb phrase. For example, consider the sentence *"skólar byrja bráðum aftur"* (schools start soon again). The correct annotation includes the two separate adverb phrases *[AdvP bráðum AdvP]* and *[AdvP aftur AdvP]*, but not the single adverb phrase *[AdvP bráðum aftur AdvP]*. The reason is that the former adverb can be moved around in the sentence (without having to move the other adverb), e.g. resulting in a sentence like *"bráðum byrja skólar aftur"*.

3.1.3 Conjunction phrases

We distinguish between two types of conjunctions: coordinating conjunctions [CP ... CP]) and subordinating conjuncitons [SCP ... SCP]). Only the following seven conjunctions are classified as coordinating conjunctions: *"og"* (and), *"en"* (but), *"eða"* (or), *"enda"* (because), *"heldur"* (but), *"ellegar"* (or), *"né"* (nor).

A conjunction phrase consists of one conjunction. The following are examples of conjunction phrases:

1. [CP og CP] (and)
2. [CP en CP] (but)
3. [SCP sem SCP] (that/who/which)
4. [SCP að SCP] (that)
5. [SCP þegar SCP] (when)

3.1.4 Interjection phrases

An interjection phrase ([InjP ... InjP]) consists of one interjection. The following are examples of interjection phrases:

1. [InjP hí InjP] (hi)
2. [InjP æ InjP] (ouch)
3. [InjP takk InjP] (thanks)
4. [InjP já InjP] (yes)

3.1.5 Adjective phrases

An adjective phrase ([AP ... AP]) consists of an adjective, optionally preceded by a modifying adverb phrase. The following are examples of adjective phrases:

1. [AP erfitt AP] (difficult)
2. [AP kalt AP] (cold)
3. [AP meira AP] (more)
4. [AP [AdvP mjög AdvP] erfitt AP] (very difficult)

5. [AP [AdvP ákaflega AdvP] fágætur AP] (extremely rare)

The first three examples show adjective phrases consisting of a single adjective, while examples no. 4-5 demonstrate adverb phrases included in adjective phrases.

3.1.6 A sequence of adjective phrases

A sequence of adjective phrases ([APs ... APs]) consists of two or more consecutive adjective phrases (optionally separated by a CP or a comma) agreeing in gender, number and case. A sequence of such phrases, typically, denote an enumeration of some kind. The following are examples of such sequences:

1. [APs [AP lágreist AP] [AP svört AP] APs] (low-rise black)
2. [APs [AP þrekinn AP] [CP og CP] [AP mikill AP] APs] (beefy and large)
3. [APs [AP stórar AP] [CP eða CP] [AP litlar AP] APs] (big or small)
4. [APs [AP gula AP] , [AP veðraða AP] APs] (yellow, weatherworn)
5. [APs [AP vörpulegur AP] , [AP skarpleitur AP] [CP og CP] [AP svipsterkur AP] APs] (pretty, sharp-featured and strong-looked)
6. [APs [AP [AdvP jafnan AdvP] grá AP] [CP eða CP] [AP skjöldótt AP] APs] (usually gray or multi-coloured)

3.1.7 Noun phrases

The structure of a noun phrase ([NP ... NP]) is the most complicated of all the phrases. In general, the unmarked word order in a noun phrase headed by a noun is an indefinite pronoun, a demonstrative pronoun/article, a numeral, an adjective phrase and a noun (and a possessive pronoun). This word order is relatively fixed with some exceptions (see below). Noun phrases can also consist of a single (personal, demonstrative, indefinite, or interrogative) pronoun.

Number, gender and case agreement holds between the words of a noun phrase. The list below shows some examples of noun phrases:

1. [NP ég NP] (I)
2. [NP sig NP] (himself/herself)
3. [NP allt NP] (all)

4. [NP þetta NP] (this)
5. [NP hvað NP] (what)
6. [NP maður NP] (man)
7. [NP 1954 NP]
8. [NP Stefán NP]
9. [NP Einar Þorgilsson NP]
10. [NP sjálfan mig NP] (myself)
11. [NP þrír fingur NP] (three fingers)
12. [NP árið 1982 NP] (year 1982)
13. [NP pabbi þinn NP] (father your)
14. [NP þetta kvöld NP] (this evening)
15. [NP [AP gömul AP] húsgögn NP] (old furniture)
16. [NP [AP nýkjörinn AP] forseti NP] (newly-elected president)
17. [NP [AP [AdvP liðlega AdvP] þrítugur AP] karlmaður NP] (a-little-more-than thirty man)
18. [NP allt þetta [AP þunga AP] vatn NP] (all this heavy water)
19. [NP enginn [AP venjulegur AP] maður NP] (no ordinary man)
20. [NP hinn [AP gagnrýni AP] efnafræðingur NP] (the critical chemist)
21. [NP þessu [AP fyrsta AP] tölublaði NP] (this first issue)
22. [NP þessi [APs [AP brúnu AP] , [AP saklausu AP] APs] augu NP] (these brown, innocent eyes)
23. [NP [APs [AP gula AP] , [AP veðraða AP] APs] múrveggnum NP] (yellow, weatherworn brick-wall)
24. [NP [APs [AP ungi AP] [CP og CP] [AP glæsilegi AP] APs] organistinn NP] (young and elegant organist)

25. [NP þess [APs [AP þriðja AP] [AP stærsta AP] APs] NP] (the third biggest)

Examples no. 1-8 show noun phrases consisting of a single word. The first two include a personal pronoun, the third an indefinite pronoun, the fourth a demonstrative pronoun, the fifth an interrogative pronoun, the sixth a common noun, the seventh a numeral and the eighth a proper noun.

Examples no. 9-14 demonstrate noun phrases comprising two words and examples no. 15-25 show adjective phrases included in noun phrases.

Some exceptions to the main word order need to be accounted for. Below we present two examples of these exceptions:

1. [NP maður einn NP] (man one)
2. [NP sinn [AP sterkasta AP] bakhjarl NP] (his strongest sponsor)

In the first example, the indefinite pronoun follows the noun (instead of preceding it), and in the second sentence the possessive pronoun precedes the adjective/noun (instead of following it).

3.1.8 A sequence of noun phrases

A sequence of noun phrases ([NPs ...NPs]) consists of two or more consecutive noun phrases (optionally separated by a CP and/or a comma) agreeing in gender, number and case. Moreover, a sequence of noun phrases can include a qualifier noun phrase which follows (or precedes) another noun phrase. A sequence of noun phrases, typically, denote an enumeration of some kind. The following are examples of noun phrase sequences:

1. [NPs [NP þrumur NP] [CP og CP] [NP eldingar NP] NPs] (thunder and lightning)
2. [NPs [NP þeim hugleiðingum NP] [CP og CP] [NP því starfi NP] NPs] (those speculations and that job)
3. [NPs [NP [AP gömul AP] húsgögn NP] , [NP [AP latneskar AP] bækur NP] [CP og CP] [NP smyrðlinga NP] NPs]
4. [NPs [NP fiskum NP] , [NP liðdýrum NP] [CP og CP] [NP spendýrum NP] NPs] (fish, arthropods and mammals)
5. [NPs [NP börn NP] [NP hans NP] [CP og CP] [NP niðjar NP] NPs] (children his and descendants)

6. [PP við [NPs [NP Lyme NP] [NP flóa NP] NPs] PP] (at Lyme bay)

The first two examples demonstrate two noun phrases separated by a coordinating conjunction phrase. The third and fourth examples show three noun phrases separated by a comma and a coordinating conjunction phrase. In the fifth example, the [NP hans NP] phrase is a genitive qualifier modifying the [NP börn NP] phrase.

The last example demonstrates a sequence of noun phrases which does not stand for an enumeration.

3.1.9 Verb phrases

Our annotation scheme subclassifies verb phrases. A finite verb phrase is labeled as [VP ... VP] and consists of a finite verb optionally followed by a sequence of adverb phrases and supine verbs. Other types of verb phrases are labeled as [VPx ... VPx] where x can have the following values:

- **i**: denoting an infinitive verb phrase
- **b**: denoting a verb phrase which demands a predicate nominative, i.e primarily a verb phrase consisting of the verb "vera" (be).
- **s**: denoting a supine verb phrase
- **p**: denoting a past participle verb phrase
- **g**: denoting a present participle verb phrase

The following are examples of verb phrases:

1. [VP hafði VP] (had)
2. [VP hafði [AdvP stundum AdvP] spjallað VP] (had sometimes talked)
3. [VP hefði [AdvP samstundis AdvP] getað ímyndað VP] (have immediately could imagined)
4. [VPi að halda VPi] (to hold)
5. [VPi að hafa VPi] (to have)
6. [VPb var VPb] (was)
7. [VPb hefur verið VPb] (has been)

8. [VPb reyndist VPb] (turned out to be)
9. [VPs staðið VPs] (stood)
10. [VPs sest VPs] (sit)
11. [VPp orðin VPp] (become)
12. [VPp kominn VPp] (arrived)
13. [VPg æpandi VPg] (screaming)
14. [VPg bölvandi VPg] (cursing)

The first example shows a finite verb phrase consisting of a single finite verb. The second and third examples demonstrate finite verbs followed by an adverb phrase and one or two supine verbs.

Examples no. 4-5 show infinite verb phrases. Examples no. 6-7 present verb phrases consisting of the verb "be", and example no. 8 includes another verb which demands a nominative complement.

Supine verb phrases are shown in examples no. 9-10. Finally, past and present participle verb phrases are demonstrated in examples no. 11-12 and no. 13-14, respectively.

3.1.10 Prepositional phrases

In general, a prepositional phrase [PP ... PP] consists of a preposition (or a MWE phrase which functions as a preposition (MWE_PP)) followed by a sequence of (one or more) noun phrases.

Case government needs to hold between the preposition and the sequence of noun phrases with the exception of an optional sequence of genitive qualifier phrases following or preceding the main noun phrases (see examples below). Furthermore, a prepositional phrase can contain an infinitive verb phrase.

Below we show examples of prepositional phrases:

1. [PP í [NP sögunni NP] PP] (in story)
2. [PP í [NP [AP skuggsælu AP] húsi NP] PP] (in shadowy house)
3. [PP á [NP [APs [AP gula AP] , [AP veðraða AP] APs] múrveggnum NP] PP] (on yellow, weatherworn brick-wall)

4. [PP í [NP sögu NP] [NP fjölskyldunnar NP] PP] (in story family's)
5. [PP [MWE_PP úti við MWE_PP] [NP sjóinn NP] PP] (out by sea)
6. [PP í [NPs [NP haustmyrkri NP] [CP og CP] [NP vetrargnauði NP] NPs] PP] (in autumn-darkness and winter-hiss)
7. [PP [MWE_PP innan um MWE_PP] [NPs [NP [AP gömul AP] húsgögn NP] , [NP [AP latneskar AP] bækur NP] [CP og CP] [NP smyrðlinga NP] NPs] PP]
8. [AP leið AP] [PP á [VPi að sitja VPi] PP] (bored on to sit)

In the first three examples, the prepositional phrases contain a single noun phrase. In the fourth example, a genitive qualifier phrase follows the main noun phrase.

A multiword expression (functioning as a preposition) precedes the noun phrase in example no. 5. Examples no. 6-7 demonstrate a preposition/multiword expression followed by a sequence of noun phrases. The last example shows an infinitive verb phrase following the preposition.

3.2 Syntactic functions

Since our constituent structure is flat, functional relations cannot be inferred from hierarchical levels. *Hence, in order to specify, for each relevant phrasal constituent, the function played within the sentence flat structures need to be augmented with explicit functional annotations* (Carroll et al. 1997).

We annotate four different types of syntactic functions: genitive qualifiers, subjects, objects/complements and temporal expressions. We use curly brackets for denoting the beginning and the end of a syntactic function (as carried out, for example, in (Megyesi and Rydin 1999)) and special function tags for labels (*QUAL, *SUBJ, *OBJ/*OBJAP/*OBJNOM/*IOBJ/*COMP, *TIMEX).

3.2.1 Genitive qualifiers

A genitive qualifier is a (sequence of) noun phrase(s), marked by the genitive case, which modifies another (usually preceding) noun phrase. The genitive qualifier is marked by {*QUAL ... *QUAL}.

Below, we show examples of such noun phrases:

1. [NP systir NP] {*QUAL [NP hennar NP] *QUAL} (sister hers)
2. [NP börn NP] {*QUAL [NP hans NP] *QUAL} (children his)

3. [NP niðurstöður NP] {*QUAL [NP þessara rannsókna NP] *QUAL} (results this research's)
4. [NP [AP nýkjörinn AP] forseti NP] {*QUAL [NP lýðveldisins NP] *QUAL} (newly-elected president republic's)
5. [PP í [NP sögu NP] {*QUAL [NP fjölskyldunnar NP] *QUAL} PP] (in story family's)
6. [PP á [NP tímum NP] {*QUAL [NPs [NP rútubíla NP] [CP og CP] [NP [AP mikilla AP] mannflutninga NP] NPs] *QUAL} PP]
7. {*QUAL [NP hennar NP] *QUAL} [NP líf NP] (her life)
8. [PP um {*QUAL [NP nokkurra ára NP] *QUAL} [NP skeið NP] PP] (over few year's period)

The first five examples demonstrate a single genitive qualifier noun phrase which modifies a preceding noun phrase. In the sixth example, a sequence of noun phrases functions as the qualifier. The last two examples show qualifier noun phrases preceding the noun phrases that they modify.

3.2.2 Subjects

Subjects in Icelandic text are (sequences of) noun phrase(s) appearing, generally, in the nominative case. Exceptions to this rule are noun phrases appearing with special finite verbs which demand subjects in the accusative or dative case. We have compiled a list of these special verbs[3].

Three possible function markers are used for subjects: *{*SUBJ> ... *SUBJ>}*, *{*SUBJ< ... *SUBJ<}* or *{*SUBJ ... *SUBJ}*. The first two tags give information about the relative position of the finite verb. **SUBJ>* means that the verb is positioned to the right of the subject, while **SUBJ<* denotes that the verb is positioned to the left of the subject. Such a relative position indicator is, for example, used in the Constraint Grammar Framework (Karlsson et al. 1995). The last tag is used when it is not clear where the accompanying verb is positioned or when the verb is missing.

Below, we show examples of subject annotations:

1. {*SUBJ> [NP ég NP] *SUBJ>} [VPb var VPb] ... (I was)

[3] Thanks to Dr. Jóhannes Gísli Jónsson, University of Iceland, for supplying the original list.

2. {*SUBJ> [NP allar óvættir NP] *SUBJ>} [SCP sem SCP] [VP bjuggu VP] ... (all ogresses which)
3. {*SUBJ> [NP systir NP] {*QUAL [NP hennar NP] *QUAL} *SUBJ>} [VPb var VPb] ... (sister hers was)
4. [VPb var VPb] {*SUBJ< [NP ég NP] *SUBJ<} ... (was I)
5. [VP kom VP] {*SUBJ< [NP [AP nýkjörinn AP] forseti NP] {*QUAL [NP lýðveldisins NP] *QUAL} *SUBJ<} ... (came newly-elected president republic's)
6. [VP kusu VP] {*SUBJ< [NPs [NP börn NP] {*QUAL [NP hans NP] *QUAL} [CP og CP] [NP niðjar NP] NPs] *SUBJ<} ... (chose children his and descendants)
7. [VP finnst VP] {*SUBJ< [NP þér NP] *SUBJ<} ... (feel-that-way you)
8. {*SUBJ [NP hauststemmning NP] *SUBJ} [PP í [NP Reykjavík NP] PP] (autumn-mood in Reykjavik)

The first three examples show a nominative case subject with the finite verb appearing to the right of it. Examples no. 4-6 demonstrate subjects for which the finite verb is positioned to the left.

Example no. 7 demonstrates a subject in the dative case – the verb "finnast" demands a dative case subject.

Finally, the last example does not have a finite verb, and thus the subject tag does not indicate relative position of the verb.

3.2.3 Objects

Our annotation scheme distinguishes between five kinds of verb complements: predicative complements (*{*COMP ... *COMP}*), direct objects (*{*OBJ ... *OBJ}*), indirect objects (*{*IOBJ ... *IOBJ}*), objects of adjectives (*{*OBJAP ... *OBJAP}*), and nominative objects (*{*OBJNOM ... *OBJNOM}*). Moreover, as is the case for subjects, "<" and ">" are used for showing the relative position of the verb.

Predicative complements are complements of verbs which demand a predicate nominative, i.e. primarily the verb *"vera"* (be), and thus appear in the nominative case. Predicative complements can be noun phrases, adjective phrases or past participle verb phrases. Predicative complements can themselves have both objects and predicative complements (see examples below).

Transitive verbs demand direct objects which can appear in any of the oblique cases. Di-transitive verbs demand both direct and indirect objects, for which, typically, the direct object is marked by the accusative case, while the indirect object is marked by the dative case (other case patterns, for direct and indirect objects, are indeed possible, e.g. dative-accusative, accusative-accusative, dative-dative and some patterns with the genitive case).

In some cases, an adjective (phrase) demands an object (see examples below).

The last type of an object, covered by our annotation scheme, is a nominative object of a verb which demands dative case subjects (see examples below).

We assume that parsers using our annotation scheme (e.g. finite-state parsers) do not resolve PP-attachment ambiguities and, thus, our scheme does not extend object noun phrases to include prepositional phrases.

Below, we show examples of object/complement annotation.

1. {*SUBJ> [NP ég NP] *SUBJ>} [VPb var VPb] {*COMP< [AP lítill AP] *COMP<} (I was small)
2. [VPb er VPb] {*SUBJ< [NP ég NP] *SUBJ<} {*COMP< [VPp fædd VPp] [CP og CP] [VPp uppalin VPp] *COMP<} ... (am I born and raised)
3. {*COMP> [AP hávaxinn AP] *COMP>} [VPb er VPb] {*SUBJ< [NP hann NP] *SUBJ<} , {*COMP< [APs [AP vörpulegur AP] , [AP skarpleitur AP] [CP og CP] [AP svipsterkur AP] APs] *COMP<} (tall is he, pretty, sharp-featured and strong-looked)
4. {*SUBJ> [NP Alís NP] *SUBJ>} [VPb var VPb] {*COMP< [VPp orðin VPp] *COMP<} {*COMP< [AP leið AP] *COMP<} (Alís had become bored)
5. {*SUBJ> [NP vagnstjórinn NP] *SUBJ>} [VP sá VP] {*OBJ< [NP mig NP] *OBJ<} (driver saw me)
6. ... [SCP sem SCP] [VP upplýsti VP] {*OBJ< {*QUAL [NP hennar NP] *QUAL} [NP líf NP] *OBJ<} (which enlightened her life)
7. ... [SCP hvorki SCP] [VPi að finna VPi] {*OBJ< [NPs [NP neinar myndir NP] [CP né CP] [NP samtöl NP] NPs] *OBJ<} (neither find any pictures nor conversations)
8. ... {*SUBJ> [NP faðmur NP] {*QUAL [NP hans NP] *QUAL} *SUBJ>} [VP umlykur VP] {*OBJ< [NP [APs [AP lágreist AP] [AP svört AP] APs] húsin NP] *OBJ<}

9. {*OBJ> [NP slíka gagnrýni NP] *OBJ>} [VP læt VP] {*SUBJ< [NP ég NP] *SUBJ<} ... (such criticism let I)

10. {*SUBJ> [NP grundin NP] *SUBJ>} [VPb var VPb] {*COMP< [VPp þakin VPp] *COMP<} {*OBJ< [NP [AP svalri AP] ábreiðu NP] *OBJ<}

11. ... [VPi að segja VPi] {*IOBJ< [NP þér NP] *IOBJ<} {*OBJ< [NP það NP] *OBJ<} (to tell you it)

12. ... [VP hefði [AdvP samstundis AdvP] getað ímyndað VP] {*IOBJ< [NP sér NP] *IOBJ<} {*OBJ< [NPs [NP eitt NP] [CP og CP] [NP annað NP] NPs] *OBJ<}

13. {*SUBJ> [NP ég NP] *SUBJ>} [VPb er VPb] {*COMP< [AP bundin AP] *COMP<} {*OBJAP< [NP Reykjavík NP] *OBJAP<} [NP [AP órjúfanlegum AP] böndum NP] (I am bound Reykjavik ...)

14. {*SUBJ> [NP honum NP] *SUBJ>} [VP fannst VP] {*OBJNOM< [NP hann NP] *OBJNOM<} [VPi sogast VPi] [PP inní PP] (He felt he suck into)

Examples no. 1-4 demonstrate predicative complements, either as adjective phrases or part participle verb phrases. The normal word order is shown in example no. 1, but variants of it are shown in examples no. 2-3. A complement of a complement is shown is example no. 4.

Examples no. 5-8 exhibit objects appearing to the right of the verb (normal word order), whereas example no. 9 shows the object appearing to the left of the verb. Example no. 10, shows a predicative complement which demands a dative object.

Examples no. 11-12 show an annotation for the objects of di-transitive verbs, i.e. indirect and direct objects appearing to the right of a verb phrase.

Finally, examples no. 13-14 show an annotation for an object of an adjective phrase, and for a nominative object of a verb which demands a dative case subject, respectively.

3.2.4 Temporal expressions

Temporal expressions in text indicate when something happened, or how long something lasted, or how often something occurs. We use *{*TIMEX ... *TIMEX}* for marking such expressions.

Below, we show examples of temporal expressions.

1. {*TIMEX [NP átta NP] *TIMEX} (eight)
2. {*TIMEX [NP árið 1982 NP] *TIMEX} (year 1982)
3. {*TIMEX [NP þetta kvöld NP] *TIMEX} (this evening)
4. {*TIMEX [NP dag einn NP] *TIMEX} (one day)

3.3 The grammar definition corpus

We have constructed a GDC, a corpus consisting of 214 sentences (3738 tokens), whose purpose is to represent the major syntactic constructions in Icelandic, in the following manner. First, we carefully selected the POS tagged sentences from the IFD corpus. Then, we used a preliminary version of our finite-state parser to automatically annotate these sentences. Finally, we checked the annotated sentences with regard to our annotation scheme and hand-corrected all the errors. Tables 1 shows the partition of the various labels for phrases and grammatical functions in our GDC.

The resulting corpus should, along with the annotation scheme itself, provide answers to questions how to analyse a given sentence in Icelandic. Furthermore, this corpus has been used to improve our parser, since we want it to be able to annotate the GDC with high accuracy.

4 Conclusion

In this paper, we have described a shallow syntactic annotation scheme – the first annotation scheme specifically designed for use by shallow parsers for Icelandic text. Our scheme comprises a set of grammatical descriptors along with examples of their use. Additionally, we have constructed a grammar definition corpus, a collection of carefully selected sentences annotated using the grammatical descriptors.

The annotation scheme has been developed as a part of a shallow parsing project.

Phrase	Frequency	%	Function	Frequency	%
MWE_CP	43	1.3%	*QUAL	103	11.9%
MWE_AdvP	21	0.6%	*SUBJ	37	4.3%
MWE_AP	1	0.0%	*SUBJ>	260	29.9%
MWE_PP	35	1.0%	*SUBJ<	100	11.5%
AdvP	231	6.7%	*OBJ>	7	0.8%
CP	183	5.3%	*OBJ<	151	17.4%
SCP	113	3.3%	*IOBJ<	8	0.9%
InjP	5	0.1%	*OBJAP>	5	0.6%
AP	291	8.5%	*OBJAP<	5	0.6%
APs	24	0.7%	*OBJNOM<	2	0.2%
NP	1308	38.1%	*COMP	11	1.3%
NPs	69	2.0%	*COMP>	140	16.1%
VP	280	8.2%	*COMP<	16	1.8%
VPi	103	3.0%	*TIMEX	24	2.8%
VPb	178	5.2%			
VPs	9	0.3%			
VPp	58	1.7%			
VPg	2	0.1%			
PP	476	13.9%			
Total:	3430	100.0%		869	100.0%

Table 1: The partition of the various labels in the GDC.

References

S. Abney. Part-of-Speech Tagging and Partial Parsing. In K. Church, S. Young, and G. Bloothooft, editors, *Corpus-Based Methods in Language and Speech*. Kluwer Academic Publishers, 1996.

J. Carroll, T. Briscoe, N. Calzolari, S. Federice, S. Montemagni, V. Pirrelli, G. Grefenstette, A. Sanfilippo, and G. Carroll. SPARKLE Work Package 1: Specification of Phrasal Parsing. Technical report, Commission of the EC, Telematics Applications Programme, Language Engineering, project LE1-2111, 1997.

EAGLES, Expert Advisory Group for Language Engineering Standards. Recommendations for the syntactic annotation of corpora. Technical report, 1996. http://www.ilc.cnr.it/EAGLES96/home.html. Accessed 15.02.2006.

G. Grefenstette. Light Parsing as Finite State Filtering. In *Proceedings of the ECAI '96 workshop on "Extended finite state models of language"*, Budapest, Hungary, 1996.

J. Hajič. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning*. Karolinum, Prague, 1998.

F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, Germany, 1995.

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

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, NJ, USA, 1994.

B. Megyesi and S. Rydin. Towards a Finite-State Parser for Swedish. In *Proceedings of the NoDaLiDa 99*, Throndheim, Norway, 1999.

J. Nivre. What kinds of trees grow in Swedish soil? A Comparison of Four Annotation Schemes for Swedish. In *Proceedings of the* 1^{st} *Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria, 2002.

J. Pind, F. Magnússon, and S. Briem. *The Icelandic Frequency Dictionary*. The Institute of Lexicography, University of Iceland, Reykjavik, Iceland, 1991.

C. Pollard and I. Sag. *Head-driven Phrase Structure Grammar*. Center for the Study of Language and Information (CSLI) Lecture Notes. Stanford University Press and University of Chicago Press, USA, 1994.

H. Þráinsson. *Íslensk setningafræði*. The Institute of Linguistics, University of Iceland, Reykjavik, Iceland, 1999.

A. Voutilainen. Designing a (Finite-State) Parsing Grammar. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*. MIT Press, 1997.

A Multiword expressions

The following tables show the list of the various multiword expressions used in our annotation scheme.

að auki	auk þess	enn einu sinni	meira segja	um leið
að minnsta kosti	á hinn bóginn	fram og aftur	mín megin	um það bil
að nokkru leyti	á ný	fyrst og fremst	nokkru sinni	vegna þess
að nýju	á stundum	hér og nú	nokkurn veginn	við og við
að sjálfsögðu	áður fyrr	hér og hvar	og svo framvegis	vinstra megin
að vísu	án þess að	hér og þar	réttu megin	þar að auki
að öðru leyti	báðum megin	hérna megin	sama hvort	þar á meðal
af og til	beggja megin	hins vegar	samt sem áður	þeim megin
aftur á bak	blátt áfram	hinum megin	sem betur fer	þess í stað
aftur á móti	eða öllu heldur	hvar sem er	sem sagt	þess vegna
aftur og aftur	ef til vill	hvers vegna	síður en svo	þín megin
alls staðar	einhvern veginn	hvort eð er	sín megin	öðru hverju
allt að því	einhvers staðar	hvort sem er	smám saman	öðru megin
allt í einu	einu sinni	hægra megin	svo og	öfugu megin
allt í lagi	ekki síst	hærra og hærra	til að mynda	
annars staðar	engan veginn	jafnt og þétt	til og frá	
annars vegar	engu að síður	meira að segja	til dæmis	

Table 2: The list of multiword expressions functioning as an adverb (MWE_AdvP).

alls konar	ferns konar	nokkurs konar	ýmiss konar
annars konar	hvers konar	sams konar	þess konar
einhvers konar	margs konar	tvenns konar	hvers kyns
eins konar	neins konar	þrenns konar	þess háttar

Table 3: The list of multiword expressions functioning as an adjective (MWE_AP).

af því að	eftir að	svo að	um leið og	því að
alveg eins	eins og	svo mikið	úr því að	því aðeins að
á meðan	enda þótt	svo mikið sem	vegna þess að	
áður en	hvorki meira né minna en	til að	þar til að	
án þess	jafnvel þótt	til þess	þar sem	
án þess að	líkt og	til þess að	þó að	

Table 4: The list of multiword expressions functioning as an adjective (MWE_CP).

aftan að	austan fyrir	fram á	fyrir neðan	inn um	neðan um
aftan af	austan í	fram eftir	fyrir norðan	inn úr	neðan úr
aftan á	austan með	fram frá	fyrir ofan	inn við	neðan við
aftan eftir	austan til	fram fyrir	fyrir sunnan	inn yfir	neðan yfir
aftan frá	austan um	fram hjá	fyrir utan	innan að	niðri á
aftan fyrir	austan úr	fram í	fyrir vestan	innan af	niðri í
aftan hjá	austan við	fram með	handan að	innan á	niður að
aftan í	austan yfir	fram til	handan af	innan eftir	niður af
aftan með	austur að	fram um	handan frá	innan frá	niður á
aftan úr	austur af	fram úr	handan fyrir	innan fyrir	niður eftir
aftan við	austur á	fram við	handan í	innan í	niður frá
aftan yfir	austur eftir	fram yfir	handan um	innan með	niður fyrir
aftur að	austur frá	framan að	handan við	innan til	niður hjá
aftur af	austur fyrir	framan af	handan yfir	innan um	niður í
aftur á	austur hjá	framan á	hér á	innan úr	niður með
aftur eftir	austur í	framan eftir	hér fyrir	innan við	niður til
aftur frá	austur með	framan frá	hér hjá	innan yfir	niður um
aftur fyrir	austur til	framan fyrir	hér í	inni á	niður úr
aftur í	austur um	framan í	hér við	inni í	niður við
aftur með	austur úr	framan með	hér undir	í gegnum	niður yfir
aftur til	austur við	framan til	inn að	í kringum	norðan að
aftur um	austur yfir	framan um	inn af	neðan að	norðan af
aftur úr	á eftir	framan úr	inn á	neðan af	norðan á
aftur við	á meðal	framan við	inn eftir	neðan á	norðan eftir
aftur yfir	á milli	framan yfir	inn frá	neðan eftir	norðan frá
austan að	á móti	fyrir aftan	inn fyrir	neðan frá	norðan fyrir
austan af	á undan	fyrir austan	inn hjá	neðan fyrir	norðan í
austan á	bak við	fyrir framan	inn í	neðan í	norðan með
austan eftir	fram að	fyrir handan	inn með	neðan með	norðan til
austan frá	fram af	fyrir innan	inn til	neðan til	norðan um

Table 5: The list of multiword expressions functioning as a preposition (MWE_PP).

norðan úr	ofan í	sunnan eftir	uppi á	út með	vestur eftir
norðan við	ofan með	sunnan frá	uppi í	út til	vestur frá
norðan yfir	ofan til	sunnan fyrir	utan að	út um	vestur fyrir
norður að	ofan um	sunnan í	utan af	út úr	vestur hjá
norður af	ofan úr	sunnan með	utan á	út við	vestur í
norður á	ofan við	sunnan til	utan eftir	út yfir	vestur með
norður eftir	ofan yfir	sunnan um	utan frá	úti á	vestur til
norður frá	suður að	sunnan úr	utan fyrir	úti í	vestur um
norður fyrir	suður af	sunnan við	utan hjá	vestan að	vestur úr
norður hjá	suður á	sunnan yfir	utan í	vestan af	vestur við
norður í	suður eftir	upp að	utan með	vestan á	vestur yfir
norður með	suður frá	upp af	utan til	vestan eftir	yfir að
norður til	suður fyrir	upp á	utan um	vestan frá	yfir af
norður um	suður hjá	upp eftir	utan úr	vestan fyrir	yfir á
norður úr	suður í	upp frá	utan við	vestan í	yfir frá
norður við	suður með	upp fyrir	utan yfir	vestan með	yfir hjá
norður yfir	suður til	upp hjá	út að	vestan til	yfir í
ofan að	suður um	upp í	út af	vestan um	yfir til
ofan af	suður úr	upp með	út á	vestan úr	yfir um
ofan á	suður við	upp til	út eftir	vestan við	yfir úr
ofan eftir	suður yfir	upp um	út frá	vestan yfir	yfir við
ofan frá	sunnan að	upp úr	út fyrir	vestur að	þrátt fyrir
ofan fyrir	sunnan af	upp við	út hjá	vestur af	
ofan hjá	sunnan á	upp yfir	út í	vestur á	

Table 6: The (continued) list of multiword expressions functioning as a preposition (MWE_PP).

REYKJAVÍK UNIVERSITY

HÁSKÓLINN Í REYKJAVÍK

School of Science and Engineering
Reykjavík University
Kringlan 1, IS-103 Reykjavík, Iceland
Tel: +354 599 6200
Fax: +354 599 6301
http://www.ru.is
ISSN 1670-5777