Specifying a Dependency Representation with a Grammar Definition Corpus

Atro Voutilainen and Krister Linden
Department of Modern Languages, University of Helsinki
first.last@helsinki.fi

Abstract

We outline the design and creation of a syntactically and morphologically annotated corpora of Finnish for use by the research community. We motivate a definitional, systematic "grammar definition corpus" as a basic step in an three-year annotation effort to help create systematically documented extensive parsebanks. The syntactic representation, consisting of a dependency structure and a basic set of dependency functions, is outlined with examples.

1. Background

This article focuses on designing a grammar definition corpus for Finnish, but first we need to say something about the purpose and context of the effort.

1.1 Treebank, Parsebank, Grammar Definition Corpus

A **Treebank** can be described as a set of sentences syntactically annotated by trained linguists. A hand-annotated Treebank is restricted in size, of high annotaation quality and consistency, and represents running text sentences and/or selected sentences illustrating various syntactic structures of the language. The PARC 700 Dependency Bank is a good example of a manually annotated Treebank, with a set of 700 text sentences annotated manually according to a form of Lexical Functional Grammar (King et al, 2003).

A **Parsebank** can be characterized by a large amount of sentences that have been mechanically annotated (with a parser), and the annotating parser has repeatedly been modified by sampling the output to correct mistakes and gradually create a better Parsebank.

In order to create a high-quality Parsebank, we need documentation and examples on the linguistic representation and its use in text analysis. A hand-annotated set of sentences is useful, but in order to approximate the structures that are used in a large corpus of text in a more comprehensive and systematic way, we need a more exhaustive and systematic set of sentences to be analysed and documented e.g. as a guideline for creating a Parsebank. We propose to use a comprehensive descriptive grammar as a source of example sentences to reach a high and systematic coverage of the syntactic structures in the language. A hand-annotated, cross-checked and documented collection of such a systematic set of sentences – in short, a **Grammar definition corpus** – should be a better approximation and guideline for annotating or parsing natural language on a large scale than a somewhat arbitrary set of sentences from txt corpora, whose relation to a comprehensive grammar is not specified.

In this paper, we outline an ongoing effort to create a Grammar definition corpus of Finnish, consisting of about 19,000 example sentences extracted from a large Finnish grammar (Hakulinen at al, 2004), and annotated according to a linguistic representation consisting of a morphological description and a dependency grammar with a basic dependency function palette.

To our knowledge, this effort if the first one based on a comprehensive, well-documented set of sentences. The closest earlier approximation to a Grammar definition corpus we know of is an English corpus, tagged and documented in the early 1990's according to a dependency-oriented representation, and consisting of about 2,000 sentences taken from a comprehensive grammar of English (Quirk et al, 1985). However, the Quirk et al grammar contains much more than the 2,000 sentences (i.e. partial coverage in the corpus), and the annotated corpus itself has not been published, though this early effort is outlined in (Voutilainen, 1997).

1.2 FINCLARIN and Parsebanking

The University of Helsinki received strategic funding for defining and creating a Research Infrastructure for the Arts and Humanities in Finland with an emphasis on various aspects of language research with the help of language technology. As part of creating a Research Infrastructure for Finnish several missing resources were identified from the start, e.g. a Finnish WordNet and a Finnish Parsebank.

As there are already various software systems in the domain of Finnish language technology, we next look at existing systems for syntactic analysis of Finnish to justify the need for the current effort.

For Finnish, there are already four different parsers owned by four different companies (Connexor, Kielikone, Sunda, Lingsoft) with slightly different syntactic approaches and with different annotation depths tailored for different applications. Most of these were created already 10-20 years ago. From a research perspective, it is problematic that none of the parsers allow free distribution of corpora that have been annotated with them. This limits their usefulness, as researchers more often find that they would like to further annotate, e.g. semantically and pragmatically, the syntactically annotated data in order to share it with others on the Internet.

In addition, the commercial parsers are available only as black boxes without the opportunity for researchers to tailor, update or improve the parsers for their own needs. A grammar definition corpus in the form of a Treebank and a subsequent Parsebank will remedy this by allowing research on new methods for automatically learning rule-based and statistical parsers from corpora as well as exploring linguistic and hybrid techniques in language modelling.

The fact that there are several divergent ways of annotating Finnish also makes it difficult to carry out linguistic product development, because choosing a new or in some respects better parser for Finnish may well mean that an application developer is faced with the laborious task of redesigning the whole linguistic application interface.

1.3 Research Infrastructure

The Parsebank serves as a test-bed when developing new language technology applications. Large amounts of syntactic structures are needed for developing information extraction, cross-language information retrieval, various machine translation technologies as well as improved grammar checkers. The intended usage points to some corpora being more useful than others for a Parsebank effort with limited resources, e.g. multi-lingual parallel corpora such as the EuroParl corpus would seem ideal, but also comparable corpora, such as Wikipedia, can be useful.

In addition, we need to make sure that the corpora and the tag sets as well as the annotated texts are publicly available with an open source license, e.g. Creative Commons, in order to avoid problems with intellectual property rights when distributing the annotated corpora.

It is also worth mentioning that another dependency corpus of Finnish, based on the Stanford Dependency scheme and articles from Wikipedia and Wikinews, is under development at University of Turku (Haverinen et al, 2009). Though the design principles of the Turku corpus differ from the present effort in several ways (e.g. corpus selection; dependency representation; use of morphology), various synergies between the corpora to be published by FIN-CLARIN and by the Turku BioNLP team remain to be explored (e.g. conversion possibilities from one linguistic representation to another).

2.Finnish Language Overview

We will not give a comprehensive overview of the Finnish language. For this, we refer to an online version of (Hakulinen et al 2004)[1] or to a succinct overview of the Finnish morphological tag set in Open Source Morphology for Finnish[2]. Instead, we will discuss some of the problems that arise when annotating Finnish

[1] http://scripta.kotus.fi/visk/etusivu.php

[2] http://home.gna.org/omorfi/omorfi/inflection.html

text syntactically, i.e. what units we wish to annotate and what morphological ambiguities we need to introduce.

2.1 Tokenization

In order to annotate a corpus, we need to decide on the granularity of the items that we intend to annotate, i.e. we need to decide on the tokens. Also in Finnish, there are multiword fixed expressions like “vähän aikaa” (some time) and “ennen kuin” (before) that do not follow the general syntactic rules of congruence and may therefore benefit from being treated as one token.

In Finnish there is a strong tendency to write compound words in one word, e.g. “yhdyssana” (compound word), but sometimes compound words have a clear multi-part structure with a blank in between, e.g. “heavy metal –henkinen” (heavy metal minded).

On the other extreme, we have enclitic words like “etten” (that I not), “miksei” (why not). In Finnish, they also inflect in all persons because the Finnish negation inflects in all persons and may take all the clitics appropriate for verbs, i.e. “ettet” (that you not), “ettei” (that he/she/it not), “ettenköhän” (that I not maybe actually [= I should probably]), etc.

The Treebank definition corpus needs to take a stand on how to tokenize multiword as well as enclitic items.

2.2 Morphology

One of the tasks of morphology is to provide the inflected words with base forms and a set of morphological tags. If the word in non-inflecting or has a deficient paradigm, we have opted for the form given by the Research Centre of Domestic Languages (KOTUS) lexicon, available at http://kaino.kotus.fi/sanat/nykysuomi.

Participles can in general be formed from all verbs, so one natural form for participles is the base form of the corresponding verb. However, some participles have clearly taken on an adjectival or nominal meaning of their own and may therefore also have the participle form as their base form. This will introduce systematic ambiguities in some cases. In Finnish there is the present participle (“-va”) , the past participle (“-nut”) , the agent participle (“-ma”) and the negation participle (“-maton”) that may introduce such ambiguities.

Derivational endings more often than not introduce a new meaning to a stem so there will be fewer mistakes by not stripping away a derivational ending. For identified derivational endings, it may still be useful to indicate the derivation, e.g. “ärsyttävästi” DRV=STI (irritatingly), even if the word is not reduced to a potential base form such as “ärsyttävä” (irritating) or “ärsyttää” (irritate).

Finnish has a rich inflectional system with thousands of forms for verbs, adjectives and nouns. Some combinations clearly have a special function and the need for reducing these to a single base form is more a question of how useful the connection with the valency or frame information of the base form is. In general, we have followed the KOTUS recommendations.

The same reasoning with regard to valency and frames also applies to newly coined derivations and it is a task for further investigations how transparent productive derivations are. From a technical point of view, a base form is simply an index to a separate semantic unit with its own syntactic behaviour. If two forms of a word have similar syntactic preferences, they may as well be reduced to the same base form, i.e. form follows function.

2.3 Syntax

Finnish syntax is characterised by (relatively) free constituent order. However, the rich Finnish morphology provides for means to express constraints on how syntactic units can be combined with each other. A parsing grammar for Finnish syntax requires extensive lexical information of valency/frame type. Such information needs to be identified from existing resources or extracted semiautomatically from large morphologically analysed corpora.

There are also some other features in Finnish grammar that need a principled classification (similar challenges occur in other languages too):

- The continuum from auxiliaries to semiauxiliaries to main verbs.
- Nominalisation (continuum from verbs to nouns), e.g. nonfinite clauses serving as pre/postmodifiers of nominals.
- Special clause types where there are no clearly idintified subjects (or rather, the subjecthood is spread over the clause)
- Conflict between surface grammatical dependency structure and semantically motivated dependency

publication. Also publication of some of the related language models will be considered.

3. Dependency Representation in Outline

In this section, we outline the dependency grammar representation used in the grammar definition corpus mostly by examples and short notes. An extensive documentation of the linguistic representation (“style sheet”) will be published separately.

Our dependency syntactic representation follows common practice in many ways. For instance, the regent of the sentence is the main predicate verb of the main clause, and the main predicate has a number of dependents (clauses or more basic elements such as noun phrases) with a nominal or an adverbial function. More simple elements, such as nominal or adverbial phrases, have their internal dependency structure, where a (usually semantic) head has a number of attributes or other modifiers.

The dependency function palette is fairly ascetic at this stage. The dependency functions for nominals include Subject, Object, Predicative and Vocative; adverbials get the Adverbial function; modifiers get one of two functions, depending on their position relative to the head: premodifying constructions are given an Attributive function tag; postmodifying constructions are given a Modifier function tag. In addition, the function palette includes Auxiliary for auxiliary verbs, Phrasal to cover phrasal verbs, Conjunct for coordination analysis, and Idiom for multiword idioms.

The present surface-syntactic function palette can be extended into a more fine-grained description at a later stage; for instance, the Adverbial function can be divided into functions such as Location, Time, Manner, Recipient and Cause. Such a semantic classification is best done in tandem with a more fine-grained lexical description (entity classification, etc).

Here are some sample analyses in tabular format. The leftmost column gives a numerical address the each token (word or punctuation mark); note that position ”0” is given as regent of the main predicate verb of the main clause. The second column from the left shows the dependency relation by indicating the position of the regent of the current word. The third column from the left shows the dependency function of the dependent. The fourth column shows the word-form itself. The fifth column shows the base form of the word (including compound boundary marker ”#”). The sixth column shows the morphological tags, e.g. word-class and inflection tags.

The quantifier ”kaikki” (all) is analysed as Attribute of the Subject noun ”peruslagerit” (basic lagers); the main predicate of the sentence ”ovat” (are) is linked (axiomatically) to ”0”, and has also another dependent, the Predicative ”samanlaisia” (similar), which has a modifying adverb ”hyvin” (very) labelled with the umbrella function Attribute.

1	2	Attribute	Kaikki	kaikki	PRON NOM

2	3	Subject	peruslagerit	peruslageri	N NOM PL
3	0	Main	ovat	olla	V ACT IND PRES PL3
4	5	Attribute	hyvin	hyvin	ADV
5	3	Predicative	samanlaisia	samanlainen	A PTV PL

Sometimes, the question arises whether to relate elements to each other on syntactic or on semantic criteria. As an example from English, consider the sentence ”I bought three litres of milk”. On syntactic criteria, the head of the object for the verb ”bought” is ”litres”, but semantically one would prefer ”milk”. Our dependency representation relates elements to each other based on semantic rather than inflectional criteria, and this has resulted in some analyses that we look at next. Note that in the following examples, base forms and morphological tags are omitted for simplicity.

Titles, roles, given names and other non-final parts of names generally are given an Attribute function rather than a nominal head function when they are followed by a suitable semantic head, e.g. surname. Hence, ”suunnittelija (planner) Marjatta [given name] are both analysed as Attribute, and both are analysed as dependents of the surname ”Nissinen”.

_6303

1	3	Attribute	Suunnittelija
2	3	Attribute	Marjatta
3	4	Subject	Nissinen
4	0	Main	asettaisi
5	6	Attribute	jokaisen
6	7	Attribute	vaatteen
7	4	Object	tekijän
8	4	Adverbial	peilin
9	8	Phrase mark	eteen
10	4	Adverbial	katsomaan
11	12	Attribute	alastonta
12	10	Adverbial	itseään
13			.

Quantifiers are analysed as Attribute of the quantified expression. For example, ”joukon” (group of) is analysed as Attribute of ”ihmisiä” (people).

_6366

1	2	Subject	Taukopaikka
2	0	Main	työllistää
3	4	Attribute	joukon
4	2	Object	ihmisiä
5			.

Likewise, ”muutaman” (a few), ”sadan” (hundred) and ”kilometrin” (kilometres) are analysed as Attribute of the adjective ”läpimittaisia” (diameter).

1	2	Attribute	Suurimmat
2	3	Attribute	tunnetut
3	4	Subject	asteroidit
4	0	Main	ovat
5	4	Adverbial	tosiaan
6	7	Attribute	vain
7	8	Attribute	muutaman
8	9	Attribute	sadan
9	10	Attribute	kilometrin
10	4	Predicative	läpimittaisia

Adpositions (prepositions and postpositions) are analysed as Phrase mark (rather than regent) of the adjacent nominal phrase. For instance, the preposition ”ennen” (before) is analysed as Phrase mark of the noun ”paluutaan” (his return). As an additional advantage, adpositional phrases receive a more similar dependency analysis with e.g. locative nominal phrases where the locative case is given morphologically (locative suffix) rather than syntactically (with an adposition). In both cases, the nominal phrase is regarded as the head category that can serve a nominal or adverbial function in the sentence.

_13217

1	2	Subject	Koivisto
2	3	Auxiliary	ei
3	4	Auxiliary	ollut
4	0	Main	saanut
5	6	Attribute	kaikkia
6	7	Attribute	syksyn
7	4	Object	saataviaan
8	9	Phrase mark	**ennen**
9	4	Adverbial	paluutaan
10	9	Modifier	kotimaahan
11	9	Modifier	joululomalle
12			.

Here is an example of a postposition: ”jälkeen” (after) is regarded as Phrase mark of the preceding nominal phrase ”tämän jutun” (this story).

_13102

1	2	Attribute	Tämän
2	4	Adverbial	jutun
3	2	Phrase mark	**jälkeen**
4	0	Main	tietääkin
5			.

Conjunctions (coordinating and subordinating) are analysed as Phrase mark for the unit that they introduce. In the case of the coordinating conjunction, e.g. ”mutta” (but), the regent of the Phrase mark function is the (head of) the following conjunct, in this case the main verb ”nähnyt” (seen) of the clause ”nähnyt kyllä olen” (seen surely have-I, i.e. I have seen). The conjunct itself is linked to the other (preceding) conjuct head, the main verb ”tunne” (know) in the clause ”en minä tunne häntä” (I don't know him).

_13175

1	3	Auxiliary	En
2	1	Subject	minä
3	0	Main	tunne
4	3	Object	häntä
5	6	Phrase mark	**mutta**
6	3	Conjunct	nähnyt
7	6	Adverbial	kyllä
8	6	Auxiliary	olen

Likewise, subordinating conjunctions are linked as Phrase marks to the head of the clause that they introduce, and the head of the clause itself is labelled with the function of the clause (e.g. object). In any case, a subordinating conjunction is sometimes optional, so the overall analysis of the clause should not change dramatically whether the optional conjunction is used or not.

A similar treatment is given also to ”se” (`it') when it is used as a formal subject or formal object: it is labelled as Phrase mark of the head of the actual subject or object, and the actual subject or object is given the appropriate function and related as dependent of the main verb.

4. Annotation process

5.1 Preprocessing

Initially the example sentences were extracted from the XML version of the electronically available online grammar. A context marker indicating where the example appeared was retained, in order to quickly be able to locate the corresponding section in the grammar description. Ideally, the context marker will let the hand-made syntax-trees be linked as illustrations of the syntactic constructions to the example sentences.

5.2 Morphology

When the samples had been expanded, they were run through two commercial parsers in order to get a first approximation for the tokenization and the morphological annotation. This was done to see how the parsers differed and also to make sure that the ground rules for the tokenization did not diverge too much from what had been found to be computationally convenient.

Most of the problems mentioned initially for the morphological annotation process arose when comparing the morphological analysis of two parsers and neither parser was considered to be superior to the other, and quite often they were both wrong – especially in cases where the example sentences represented a Finnish dialect or informal spoken language. This strengthened our conviction that a public and well-documented definition corpus for annotating Finnish is extremely important for enabling linguistic development and promoting the adoption of language technology in research prototypes as well as applications.

5.3 Syntax

The manual tagging of the syntactic dependencies and functions was done by three linguists with background in Finnish linguistics working on separate sections of the grammar definition corpus, after a training period. The data for annotation was given in a spreadsheet format, with the columns for dependency relation and dependency function to be populated by the annotators.

During the annotation period, 1-2 meetings were arranged each week to discuss and resolve e.g. borderline cases between different analyses. As a result of the discussions, the documentation of the dependency syntactic representation was extended and made more specific. Problematic cases and misanalyses were detected by the annotators when checking their own annotations; additional cases and inconsistencies were found as a result of cross-checks between the annotators. In case of genuinely problematic cases, the annotators were instructed not to force an arbitrary analysis, but to leave the problematic part of the sentence unanalysed, and to discuss them in the weekly meetings. Manually providing dependency syntactic functions and dependency relations for the 19,000 example sentences took approximately 5 person months.

6. Further Work

After the first annotation of the corpus, automatic consistency checks are made to flag remaining problematic cases for expert revision. The morphology needs to be realigned with the syntactic analysis decisions. This may necessitate redefinition of some morphological categories.

The current linguistic representation is coarse-grained about several areas of grammatical description (e.g. adverbial functions), but the present representation is designed to support more fine-grained functional analysis in areas such as description of modality and negation; entity classification and normalisation; sentence type analysis; anaphora and coreference resolution; even some degree of synonymy resolution (using such recent resources as Finnish WordNet).

Acknowledgements

Prof. Maria Vilkuna from KOTUS provided us with valuable expert consultation on Finnish morphology and grammar at various stages of the project. Mirja Nissinen and Sam Hardwick preprocessed the grammar definition corpus, provided it with a morphological analysis, corrected many of the erroneous analyses, and generally prepared it for the dependency syntactic analysis. Mikaela Klami, Satu Leisko-Järvinen and Tanja Purtonen have manually applied the dependency syntax to the corpus, supported the documentation effort, and given valuable feedback and ideas on the analysis solutions used in this annotation effort.

References

Hakulinen, A.; Vilkuna, M.; Korhonen, R.; Koivisto, V.; Heinonen, T.; Alho, I. *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura. 2004.

Haverinen, K.; Ginter, F.; Laippala, V.; Viljanen, T. & Salakoski, T.: Dependency Annotation of Wikipedia: First Steps towards a Finnish Treebank. *Proceedings of The Eighth International Workshop on Treebanks and Linguistic Theories (TLT8*). 2009. http://bionlp.utu.fi/sites/default/files/haverinen-et-al-2009.pdf..

King, Tracy; Crouch, R.; Rietzler, S.; Dalrymple, M.; Kaplan, R.M. The PARC 700 Dependency Bank. In *Proceedings of the 4th Interntional Workshop on Linguistically Interpretd Corpora, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest. 2003.

Atro Voutilainen. Designing a (Finite-State) Parsing Grammar. In Emmanuel Roche and Yves Schabes (editors): *Finite-State Language Processing*. The MIT Press. Cambridge, Massachusetts; London, England. Pages 283-310. 1997.

Turku dependency treebank 2010: http://bionlp.utu.fi/fintreebank-finnish.html..

OMORFI: http://gna.org/projects/omorfi
http://www.ling.helsinki.fi/kieliteknologia/tutkimus/omor/index.shtml