Specifying Treebanks, Outsourcing Parsebanks: FinnTreeBank 3

Atro Voutilainen & Kristiina Muhonen & Tanja Purtonen & Krister Lindén

University of Helsinki
Department of Modern Languages
FIN-CLARIN
{first.last}@helsinki.fi

Abstract

Corpus-based treebank annotation is known to result in incomplete coverage of mid- and low-frequency linguistic constructions: the linguistic representation and corpus annotation quality are sometimes suboptimal. Large descriptive grammars cover also many mid- and low-frequency constructions. We argue for use of large descriptive grammars and their sample sentences as a basis for specifying higher-coverage grammatical representations. We present an sample case from an ongoing project (FIN-CLARIN FinnTreeBank) where an grammatical representation is documented as an annotator's manual alongside manual annotation of sample sentences extracted from a large descriptive grammar of Finnish. We outline the linguistic representation (morphology and dependency syntax) for Finnish, and show how the resulting 'Grammar Definition Corpus' and the documentation is used as a task specification for an external subcontractor for building a parser engine for use in morphological and dependency syntactic analysis of large volumes of Finnish for parsebanking purposes. The resulting corpus, FinnTreeBank 3, is due for release in June 2012, and will contain tens of millions of words from publicly available corpora of Finnish with automatic morphological and dependency syntactic analysis, for use in research on the corpus linguistics and language engineering.

Keywords: treebank, outsourcing, Finnish

1. Introduction

Researchers and developers in academia and industry need a facility that enables them to easily download as well as disseminate empirical data for testing, validating and refining scientific hypotheses, models and claims. To answer these needs, there is an ongoing project at the University of Helsinki and its collaborators as part of the CLARIN and META-NET frameworks to create a user-friendly web service for researchers and developers in Finland and other countries. As part of this effort, there is ongoing work to create an extensive and easily available dependency syntactic treebank and parsebank for the Finnish language.

In this paper, we present FinnTreeBank 3, a treebank and parsebank for Finnish, and focus on its design and development. First, we describe our method of specifying the core linguistic representation with descriptive grammars (rather than purely with samples of naturally occurring text). In Section 3, we outline the linguistic representation used in describing Finnish morphology and syntax. In Section 4, we describe one use of a systematically specified linguistic representation and grammar definition corpus: as a task specification for a subcontractor to create a parser engine and a parsebank for the language resource service. We conclude with a look at future developments.

2. Building Treebanks with Descriptive Grammars

Most large-scale treebanks are based on running naturally occurring texts from different genres (e.g. news, fiction). During the treebanking or annotation process, the annotator's manual (documentation of the linguistic descriptors and their application guidelines) is built and updated to account for emerging phenomena (linguistic structures, borderline cases, etc.). As a result, the treebank, the annotation scheme and its documentation is likely to account well for high frequency phenomena that occur in the object language.

An attested weakness of this purely corpus-based approach is that many low-frequency constructions are likely to emerge as an undocumented and even problematic case so late during the treebanking process that satisfactorily accounting for the encountered construction is no longer feasible: revising the existing annotation to make it agree with the updated annotation scheme is probably too resource-consuming for the project at an advanced stage. As an example, at a recent CLARA treebanking course (Prague, December 2010), treebanking projects presented cases like the English "the + Comparative .. the + Comparative" as problematic for the annotation schemes developed during the annotation projects. – If we look at any large-coverage descriptive English grammar, we find careful accounts and examples of mid- and low-frequency constructions like "the + Comparative .. the + Comparative".

To create a high-quality parser and treebank, documentation and examples on the linguistic representation and its use in text analysis are needed. To approximate also less frequent structures used in a large corpus of text in a comprehensive and systematic way, we need a maximally exhaustive and systematic set of sentences to be analyzed and documented e.g. as a guideline for creating a parsebank. We used a comprehensive descriptive grammar as a source of example sentences to reach a high coverage of the syntactic structures in the language. A hand-annotated, cross-checked and documented collection of such a systematic set of sentences – a grammar definition corpus – is a workable initial approximation and guideline for annotating or parsing natural language on a large scale. The initial definitional sentence corpus can be extended with new data when

leaks in the grammar/corpus coverage become evident e.g. on the basis of double-blind annotations (Voutilainen and Purtonen, 2011).

A corpus containing only example sentences from a descriptive grammar obviously does not contain linguistic constructions with the frequencies that they have in large samples of naturally occurring text. Quantitative studies also need text frequencies; hence we also need large annotated samples of naturally occurring text. The main role of the grammar definition corpus is to support the specification of a large-coverage grammatical representation or annotation scheme. The combined annotation scheme and grammar definition corpus itself can be used as a resouce for well-documented annotation of large text corpora for use in (quantitative) corpus studies.

Our Finnish grammar definition corpus consists of about 19,000 example utterances extracted from a comprehensive Finnish grammar (Hakulinen et al., 2004a). It is manually annotated according to a a dependency grammar with a basic dependency function palette. This initial manually annotated corpus is called FinnTreeBank 1. Together with manually annotated smallish samples of natural texts (fiction, Finnish Wikipedia), it is the basis for an upcoming extended version, FinnTreeBank 2. The consistency and applicability of the dependency syntactic representation coded in FinnTreeBank 1 has been empirically evaluated in a pilot experiment by Voutilainen and Purtonen (2011) with results that support the practicality of the using a grammar definition corpus.

3. Linguistic Representation

Regarding syntactically annotated large Finnish corpora, there is also an ongoing effort at the University of Turku on manual annotation of Finnish corpora according to a related dependency syntactic representation (Haverinen et al., 2010). The Turku Dependency Treebank, based on the Stanford dependency scheme, currently covers about 4,300 sentences.

While the general trend in treebank annotation has been to use running text (e.g. newspaper articles) as the empirical basis of linguistic specification, our project is to our knowledge the first one in its use of a large descriptive grammar (Hakulinen et al., 2004a) as the empirical basis of the linguistic representation and treebank.

FinnTreebank 1 is a collection of example sentences taken from the online version of the descriptive grammar, VISK (Hakulinen et al., 2004b). In the treebank, the VISK sentences are presented in a tabular, CONLL-X standard-conforming format. The CONLL-X standard has ten data types (fields), of which seven are utilized in the analysis of the VISK corpus. Table 1 portrays the dependency syntactic representation for the sentence *Haluaisin kuitenkin esittää yhden pyynnön.*, glossed in Example (1) below.

(1) Haluaisin kuitenkin esittää yhden pyynnön.
want however make one wish
I would, however, want to make a wish.

In the CONLL-X format of FinnTreeBank 1, all word forms and punctuation marks are presented on a separate line. Each word has a numerical address within the sentence (column 1). The next column from the left is the actual word form, followed by its base form in column 3. The morphological description is given in both the short, coarse-grained manner (column 4), and a fine-grained analysis (column 5).

The dependency relations (dependent–regent relations), are marked in column 7 by indicating the governing word (regent) using the sentence-internal numerical address of column 1. For instance, the word form *esittää (make)* is governed by its regent at position 1, *haluaisin (want)*. The nucleus of the sentence is usually the main verb of the main clause. In Table 1, the verb *haluaisin (want)*, takes the (non-existent) regent at position 0.

The dependency functions of the word forms are presented in column 8. In Example (1), *pyynnön (wish)* functions as an object of its regent *esittää*. Columns 6, 9 and 10 of the CONLL-X standard are not used in FinnTreeBank; the unused fields are marked with an underscore, (_).

3.1. Morphology

The morphological and lexical representation of FinnTreeBank covers the rich inflectional and compounding morphology of Finnish. Also basic derivational morphology of Finnish is covered (Pirinen, 2008). Morphology is analyzed using the HFST tools and OMorFi analyzer of the Helsinki HFST team (Lindén et al., 2009).

The morphological analysis used in FinnTreeBank is based on the descriptive grammar (Hakulinen et al., 2004a) as much as is possible. Since the descriptive grammar does not offer an unambiguous solution to all morphological questions, we need to think beyond the descriptive grammar and come up with a reasonable way of modeling the phenomena within our annotation scheme. Morphological phenomena that have more than one obvious analysis is done in a symmetrical, balanced, semantically motivated manner.

Implementing a symmetrical and balanced annotation model means that if a phenomenon can be analyzed in several ways, we follow the same ground rules in solving them. An example of this is words that possess traits of different word classes, like participles acting both as adjectives and as verbs. In addition to participles, a specific example of such word classes are infinitives. In FinnTreeBank, a word inflected in the third infinitive that is used as an adverbial of manner can be seen as a verb or an adverb on the morphological level.

Some of the third infinitives are at least partially lexicalized as adverbs. In particular, in contexts where the third infinitives give information about modality, they are seen as adverbs or particles. These contexts can be distinguished from contexts where they act as verbs: As adverbs or particles, third infinitives do not have extensions, as can be seen in Table 2 .

In Table 2, the third infinitive *epäilemättä (undoubtedly)* is analyzed as an adverb because it does not have any extensions. The infinitive modifies the main verb of the sentence, *on (is)*.

The analysis of third infinitives changes if the word takes extensions. This is demonstrated in Table 3.

1	Haluaisin	haluta	Verb	V Act Cond Sg1	_	0	main	_	_
2	kuitenkin	kuitenkin	Adverb	Adv	_	1	adverbial	_	_
3	esittää	esittää	Verb	V Act Inf1 Sg Lat	_	1	object	_	_
4	yhden	yksi	Num	Num Card Sg Gen	_	5	attribute	_	_
5	pyynnön	pyyntö	Noun	N Sg Gen	_	3	object	_	_
6	.	.	Punct	Punct Sent	_			_	_

Table 1: The CONLL-X format used in FinnTreeBank 1

#	WORD	TRANS	M	R	F
1	Se	it	Pron Dem Sg Nom	2	Subj
2	on	is	V Act Ind Prs Sg3	0	Main
3	epäilemättä	undoubtedly	Adv	2	Advl
4	vaikeaa	hard	A Pos Sg Par	2	Scomp
5	.		Punct Sent		
	It is undoubtedly hard.				

Table 2: Third infinitive *epäilemättä* as an adverb

#	WORD	TRANS	M	R	F
1	Hän	she	Pron Pers Sg Nom	2	Subj
2	uskoi	believed	V Act Ind Prt Sg3	0	Main
3	kaiken	everything	Pron Qnt Sg Gen	2	Obj
4	mitään	nothing	Pron Qnt Sg Par	5	Obj
5	epäilemättä	doubt	V Act Inf3 Sg Abe	2	Advl
6	.		Punct Sent		
	She believed everything without a doubt.				

Table 3: Third infinitive *epäilemättä* as a verb

In Table 3, the potential adverb *epäilemättä* is seen as a verb because of its verbal extension, object *mitään*.

In sum, when lexicalized third infinitives have valency-based extensions, they are seen as verbs on the morphological level. When they occur without an extension, they are adverbs on the morphological level. Considering whether or not a word takes extensions when making the annotation decision for borderline cases is one of the main principles we follow in the annotation scheme of FinnTreeBank. The same principle, demonstrated in Tables 3 and 2, is used e.g. when annotating verb-to-noun derivations and participial constructions.

3.2. The Dependency Syntactic Representation

The grammatical model and the syntactic annotation schemes are based on dependency syntax (Tesnière, 1980). From this follows that words are linked together with unidirectional two-term relations (dependencies) which are labeled with their dependency functions.

Our dependency syntactic representation follows common practice; for instance, the head 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. The dependents may have their own internal dependency structure, e.g. a subordinate clause may have its own subject and object etc.

The dependency function palette is fairly ascetic and contains only 15 functions. However, it is extensible into a more semantic representation, e.g. adverbial subclasses such as location, time and manner.

Our dependency representation relates elements to each other based on semantic rather than inflectional criteria. Hence, our analysis gives a dependent role to categories such as conjunctions, prepositions, postpositions, auxiliaries, determiners, quantifiers, attributes and formal elements (formal subject, formal object, etc.).

The syntactic analysis is shallow and non-projective. A shallow syntactic analysis means that no (empty) word-like categories are postulated, so the analysis is based on word forms that actually exist in the sentence. This means that e.g. no missing verb is postulated in an elliptical clause. Moreover, the dependency relations used in the language model are non-projective, making it possible for the model to capture long-distance dependencies. This is crucial in languages with free constituent order.

4. Parsebanking by Outsourcing: FinnTreeBank 3

One estimated advantage of building and documenting a treebank with systematically collected example sentences from a large descriptive grammar is that the resulting corpus and documentation can be used as a task specification for an external Human Language Technology (HLT) supplier. We tested this idea by outsourcing morphological and dependency syntactic annotation of two large Finnish corpora (Europarl and JRC-Acquis) to a commercial HLT provider. We discuss the outsourcing process in more detail in Section 4.1.

When evaluating the subcontractor's annotation, we detected some unwanted annotations. These annotations were caused by the phenomena not being specified strictly enough in the annotation manual. These phenomena had been overlooked in the manual writing phase because the annotators from the FinnTreeBank team annotated the phenomena similarly in initial double-blind tests (see e.g. (Voutilainen and Purtonen, 2011)). Thus, outsourcing the parsebank creation revealed shortcomings of the original linguistic specification which the FinnTreeBank team had

not detected at first, but which can later be fixed and completed. In addition to other benefits of outsourcing, we can regard outsourcing as a method for improving the manual, the linguistic specification of the annotation scheme. This is discussed in Section 4.2.

4.1. Outsourcing

Outsourcing the parsebank creation was executed on the basis of a request for quotation to four Finnish HLT companies. Each proposal contained a guarantee for certain accuracy figures related to morphology and dependency syntax (lemma, word-class, dependency function, dependency relation). The development team of the selected supplier was provided with hands-on training on the use of the linguistic representation, and a chance to communicate with the FIN-CLARIN research team on the linguistic specification. Additional feedback to the supplier was provided on the basis of intermediate deliveries (samples of automatically annotated corpora) that were evaluated by the FIN-CLARIN research team.

The annotation of the Europarl and JRC-Acquis corpora was based on a combination of a parser package (mate-tools by Bohnet (2010)) and a commercial rule-based morphological analyzer. Most of the subcontractor development was carried out during 2011, and the resulting FinnTreeBank 3 (consisting of previously released manually annotated treebanks, automatically parsed Europarl and JRC-Acquis parsebanks in tabular CONLL-X format and documentation of the grammatical representation) became available during the spring of 2012.

4.2. Improving the Manual

Before outsourcing the parsebank creation, we created the annotation manual describes annotation principles and explains example annotations for different linguistic phenomena. The manual was written during the creation of the first manually annotated treebank, FinnTreeBank 1. Writing the annotation manual simultaneously with the treebank annotation enables instant testing of the annotation decisions. The scheme could thus be modified based on its usability during the development phase.

After publishing FinnTreeBank 1 and the annotation manual for it, we tested the annotation scheme using the double-blind method (Voutilainen and Purtonen, 2011). We analyzed the errors caused by incomplete specification of the phenomena in the manual and created the annotation principles for the most frequent unspecified phenomena. We used running text instead of the example sentences of the descriptive grammar, and the double-blind test corpora contained many expressions, e.g. dates and idioms, which were not specified fully in the manual. Based on the test, we improved our manual and added e.g. annotation principles for numeric expressions. The following example of a date expression (Table 4) demonstrates these main principles.

In the annotation scheme for dates described in Table 4 we follow the same principles as in other numeric expressions: The inflected word of the NP is seen as a dependent of the verb. In Table 4, the inflected word is *joulukuuta (December)*. Other words of the NP are dependents of this inflected word.

#	WORD	TRANS	M	R	F
1	Halonen	Halonen	N Prop Sg Nom	2	Subj
2	syntyi	was born	V Act Ind Prt Sg3	0	Main
3	24.	24.	Num Ord Sg Nom	4	Attr
4	joulukuuta	December	N Sg Par	2	Advl
5	1943	1943	Num Card Sg Nom	4	Mod
6	.	.	Punct Sent		
	Halonen was born on December 24, 1943.				

Table 4: Annotation scheme for dates

Underspecified phenomena like date expressions were spotted and reported in the manual before outsourcing. The reason for the phenomena being undetected at first is that the annotators of FinnTreeBank 1 shared the same linguistic background and annotated structures similarly without defining them specifically. Therefore, some structures remained unconsciously underspecified.

In addition to other benefits, the outsourced treebank can be used as means to reveal underspecified phenomena and to complete the manual. The wide coverage of the manual improves the usability of the treebank and is important especially from the evaluation's point of view: Before we can publish any correctness rates, the correct analysis must be defined clearly.

The correctness of the outsourced treebank was evaluated using cross-checking and double-blind tests. The differences were analyzed and cases where the differences were caused by the manual being incomplete were reported. After the evaluation of the subcontractor's annotation, we noticed that the manual did not specify e.g. the structure *mikä X tahansa (whichever X)*, shown in Table 5.

#	WORD	TRANS	M	R	F
1	Valitse	pick	V Act Imprt Sg2	0	Main
2	näistä	these+elative	Pron Dem Pl Ela	1	Advl
3	mikä	which	Pron Qnt Sg Nom	5	Attr
4	kortti	card	N Sg Nom	1	Obj
5	tahansa	(which)ever	Adv	4	Mod
6	.	.			
	Pick any of these card.				

Table 5: a

The annotation scheme for the structure described in Table 5 was problematic to define because the structure does not have a fixed word order and can be elliptical. The word *mikä (which(ever))* is seen as dependent of the word *tahansa ((which)ever)*, and not as a dependent of the noun because the noun is not necessarily realized.

In addition to adding annotation schemes for very specific structures like the one described in Table 5, we have created more general annotation principles. The purpose of the more general principles is that they can be applied when a new, unspecified structure occurs.

5. Conclusions and Future Developments

In this paper, we have argued for use of descriptive grammars and their sample sentences as an initial approximation of the structures that occur in natural language texts, to maximise the coverage of the grammatical representation or annotation scheme in annotation of large corpora of naturally occurring texts. One of the uses of the resulting Grammar Definition Corpus (with annotated sample sentences) and the accompanying documentation (annotator's manual) is as a specification for a parser's language model. We have described a subcontracting process, where the subcontractor built a language model and a combined morphological analyser and parser of Finnish for annotating large volumes of naturally occurring text for inclusion in Finnish parsebanks. The first version of the resulting parsebank, FinnTreeBank 3, is due for release in June 20111, and it will contain Europarl and JRC-Aquis texts with morphological and dependency syntactic analysis in CONLL-X format.

There is also need for further developments to support Finnish parsebanking and corpus linguistics in the near and more distant future:

- development and integration of a web interface for pattern extraction and visualisation.
- refinement of extraction and reporting based on user needs.
- development of alternative or complementary language models for parsing text with higher accuracy. In addition to the subcontracting process resulting in a hybrid parsing model (linguistic morphology, statistical parsing), we have also started work on a fully linguistic model based on finite state morphology and constraint grammars.
- development of semiautomatic methods for treebanking to support efficient annotation of 'medium-large' volumes of text (millions of words) with a accuracy levels that cannot be reached automatically.
- design and development of annotated corpora with a more informative linguisic analysis. For instance, the present linguistic representation can be extended with recognition and classification of names and with a semantically oriented functional account of syntactic relations.

Acknowledgements

The ongoing project has been funded via CLARIN, FIN-CLARIN, FIN-CLARIN-CONTENT and META-NORD by EU, University of Helsinki, and the Academy of Finland. We would like to thank the three anonymous reviewers for their constructive comments.

6. References

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In optional optional, editor, *The 23rd International Conference on Computational Linguistics (COLING 2010)*, volume optional of *optional*, page optional, Beijing, China, optional. optional. optional.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho. 2004a. *Iso suomen kielioppi*. Suomalaisen Kirjallisuuden Seura, Helsinki. ISBN: 951-746-557-2.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho. 2004b. Ison suomen kieliopin verkkoversio: määritelmät. Suomalaisen Kirjallisuuden Seura. http://kaino.kotus.fi/cgi-bin/visktermit/visktermit.cgi.

Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski. 2010. Treebanking Finnish. In Markus Dickinson, Kaili Müürisep, and Marco Passarotti, editors, *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 79–90.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology – an efficient open-source package for construction of morphological analyzers. In *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology*.

Tommi Pirinen. 2008. Suomen kielen äärellistilainen automaattinen morfologinen analyysi avoimen lähdekoodin menetelmin. Master's thesis, Helsingin yliopisto.

Lucien Tesnière. 1980. *Grundzüge der strukturalen Syntax*. Klett-Cotta, Stuttgart. ISBN: 3-12-911790-3.

Atro Voutilainen and Tanja Purtonen. 2011. A double-blind experiment on interannotator agreement: The case of dependency syntax and Finnish. In *NODALIDA 2011 Conference Proceedings*, pages 319–322.