Designing a dependency representation and grammar definition corpus for Finnish

ATRO VOUTILAINEN, KRISTER LINDÉN, TANJA PURTONEN

Department of Modern Languages, University of Helsinki

atro.voutilainen@helsinki.fi, krister.linden@helsinki.fi, tanja.purtonen@helsinki.fi

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 first step in a three-year annotation effort to help create higher-quality, better-documented extensive parsebanks at a later stage. The syntactic representation, consisting of a dependency structure and a basic set of dependency functions, is outlined with examples. Reference is made to double-blind annotation experiments to measure the applicability of the new grammar definition corpus methodology.

Parsebank, grammar definition corpus, dependency grammar

Presentamos el primer diseño y creación de un corpus del finlandés anotado sintáctica y morfológicamente para su uso por la comunidad científica. En este trabajo se motiva un "corpus de definición gramatical" sistemático y que servirá como base para un proyecto de anotación de tres años, como ayuda para la creación de corpus anotados sintácticamente (treebanks o parsebanks) amplios, de mejor calidad y mejor documentados en una fase subsiguiente. La representación sintáctica, consistente en una estructura de dependencias y un conjunto básico de funciones de dependencia, es presentada con ejemplos. En este trabajo se hace referencia a los experimentos de anotación doblemente ciegos (double-blind) para medir la aplicabilidad de la nueva metodología para el corpus de definición gramatical.

1. BACKGROUND

This paper outlines the first main step - motivation and design of a grammar definition corpus - in a multiyear project at University of Helsinki (as part of the pan-European CLARIN research infrastructure effort) to provide (i) open-source morphological and dependency syntactic language models and analysers for the Finnish language and (ii) publicly available morphologically and dependency syntactically annotated large text corpora of Finnish (e.g. Finnish Wikipedia and EuroParl corpora) for R&D uses in Finland and other countries.

More specifically, we outline an effort to create a **grammar definition corpus** and related documentation of linguistic descriptors ("stylesheet") of Finnish. This corpus consists of 19,000 example sentences extracted from a comprehensive descriptive Finnish grammar (Hakulinen, Vilkuna, Korhonen, Koivisto, Heinonen & Alho, 2004), and annotated according to a linguistic representation (a morphological and dependency syntactic grammar with a basic dependency function palette). To our knowledge, this effort if the first one based on a comprehensive, systematic set of sentences illustrating the syntactic structures of a natural language in considerable depth. This grammar definition corpus will be used as a basis for creating and documenting (i) formal language models and parsers for use in automatic corpus annotation and (ii) large syntactically annotated text corpora for R&D related to the Finnish language.

The structure of this paper is as follows. Section 2 discusses the terms "treebank", "parsebank" and "grammar definition corpus". Section 3 outlines descriptive solutions related to Finnish language analysis. Section 4 focuses on the dependency syntactic representation used in the grammar definition corpus. Section 5 tells about the work process and deliverables.

2. 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 annotation 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, Crouch, Rietzler, Dalrymple & Kaplan, 2003). Far larger annotated resources of English are documented in (Cinková, Toman, Hajič, Čermáková, Klimeš, Mladová, Šindlerová, Tomšů & Žabokrtský, 2009; Marcus, Santorini & Marcinkiewicz, 2004). Additionally, Wikipedia ("Treebank") lists a large number of treebank projects for many languages.

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 use a large 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* – serves as an inventory of high and low frequency syntactic constructions in the language.

However, sample sentences in a descriptive grammar usually are kept as simple and short as is convenient for illustrating the grammatical construction in point. To start approximating the variation possibilities within each grammatical construction, additional running-text corpora from different genres are needed for annotation – but following the guidelines set at the definitional phase.

3. FINNISH IN OUTLINE

Morphology. Finnish has a rich inflectional system with thousands of forms for each verb, adjective and noun. 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.

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 descriptive grammar (Hakulinen *et al.*, 2004) .

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. Ambiguities between lexicalised and systematic analyses can be resolved in lexicalised parsing grammars as documented in Voutilainen (2003), so emergence of such ambiguities is not considered problematic.

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 is still 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).

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.

Syntax. Finnish syntax is characterised by (relatively) free constituent order. 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 from large morphologically analysed corpora.

There are also some other features in Finnish grammar that need a principled (or at least operational) classification (similar challenges occur in other languages too): (i) analysis of so-called special clause types (where the potential subject has an untypical case); (ii) continuum from auxiliaries to semiauxiliaries to main verbs (a similar continuum exists in other languages too, e.g. English (Quirk & al 1985: 136-147); (iii) nominalisation (continuum from verbs to nouns). The grammar definition corpus drawn from Hakulinen *et al*. illustrates continua such as these with numerous well-ordered example sentences, which helps make a systematic categorisation.

4. 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. A larger 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. In our representation, grammatical markers (such as determiners, conjunctions, auxiliaries and adpositions) are described as dependents (with an attributive or phrase marker or auxialiary function); as a result, semantically "heavier" words get a head status in dependency analyses. In this respect, our representation

follows that used in the Prague Dependency Treebank (while e.g. the Danish Dependency Treebank follows almost the opposite policy of granting grammatical categories a head status).

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 (attr) of the Subject (subj) noun *peruslagerit* (basic lagers); the main predicate verb of the sentence *ovat* (are) is linked (axiomatically) to ”0”, and has also another dependent, the Predicative (pred) *samanlaisia* (similar), which has a modifying adverb *hyvin* (very) labelled as Attribute.

1	2	attr	Kaikki	kaikki	all	PRON NOM PL
2	3	subj	peruslagerit	peruslager	basic-lager	N NOM PL
3	0	main	ovat	olla	be	V ACT IND PRES PL3
4	5	attr	hyvin	hyvin	very	ADV
5	3	pred	samanlaisia	samanlainen	similar	A PTV PL

Table 1. “All basic lagers are very similar.”

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. Also quantifiers are analysed as Attribute of the quantified expression. For example, *joukon* (group of) is analysed as Attribute of *ihmisiä* (people).

1	2	subj	Taukopaikka	tauko#paikka	rest-place	N NOM SG
2	0	main	työllistää	työllistää	employ	V ACT IND PRES SG3
3	4	attr	joukon	joukko	group-of	N GEN SG
4	2	obj	ihmisiä	ihminen	people	N PTV PL

Table 2. “The resing place employs a group of people.”

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.

1	2	subj	Koivisto	Koivisto	Koivisto	N NOM SG
2	3	aux	ei	ei	not	NEG
3	4	aux	ollut	olla	have	V ACT SG3
4	0	main	saanut	saada	receive	V ACT PCP PAST SG

5	6	attr	kaikkia	kaikki	all	PRON PTV PL
6	4	obj	saataviaan	saatava	receivable	N PTV PL POSS
7	8	pmark	ennen	ennen	before	PREP
8	4	advl	paluutaan	paluu	return	N PTV SG POSS

Table 3. “Koivisto had not received all of his receivables before his return.”

Also 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. The conjunct itself is linked to the other (preceding) conjuct head.

5. ANNOTATION AND DELIVERABLES

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 week's 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 weekly meetings were arranged to discuss and resolve e.g. borderline cases between different analyses. In addition, the annotators cross-checked each other's output to detect possible interannotator inconsistencies. The highest consistency would probably have been reached using double/triple-blind method combined with negotiations (Voutilainen, 1999), but this method was not used due to resource and time limitations.

As a result of the discussions, the documentation of the dependency syntactic representation was extended and made more specific. Problematic cases and outright misanalyses were often detected by the annotators when checking their own annotations; additional cases and inconsistencies were found as a result of daily 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 bring it to the weekly meetings. The work on syntactically annotating the grammar definition corpus of the 19,000 grammar sentences by hand took approximately 5 person months.

The 19,000-sentence grammar definition corpus and documentation has been published (contct details to be provided); additional corrected versions will follow through 2011-2012.

A limited amount of running text representing different genres and taken from various public sources has also been annotated manually according to the dependency syntax specification resulting from the grammar definition phase. This step provides additional high-quality annotated corpus for researchers (e.g. to serve as additional learning and testing material for building language models for rule-based and statistical parsers). In addition, this step will help experiment with the usability of the developed grammar scheme in the analysis of real-world text; in terms of coverage and consistency, for instance. The manually annotated corpus will be published during 2011.

Initial experiments on interannotator agreement using the double-blind method and negotiations with limited data (three texts from different genres amounting to over 200 sentences) have been carried out to assess the pros and cons of using a systematic set of example sentences from a descriptive grammar as the initial data in a treebank (anonymous citation, to be provided). The main observations were that after negotiations, the interjudge agreement at word level (labelled dependency relations) was close to 99%. During the negotiations it was found that also complex syntactic phenomena, including various mid or low frequency special sentence types, were generally annotated quite consistently among the annotators, even before the negotiation phase took place. This supported the hypothesis that a grammar definition corpus would cover a high number of syntactic constructions in the language, and the resulting treebank and documentation should guide annotation of sentences containing these syntactic phenomena.

During the experiments it was also found that annotations were unsystematic mostly in expressions including numerals and referring to temporal or areal phenomena, which are typically poorly covered (maybe as linguistically “uninteresting phenomena”) in traditional descriptive grammars. In the case of such semi-structured phenomena, the need to negotiate a

consistent analysis to be documented in the annotator's manual and exemplified in the grammar definition corpus, became evident.

6. WORK TO DO

The ongoing project will deliver also large corpora from public sources (such as the Finnish EuroParl corpus) analysed automatically following the dependency syntax specification described above. The automatic analysis (or alternative analyses) will result from language models and parsers made according to the grammar definition corpus and its documentation. The accuracy of the automatic analysis will be lower than is the case with the manually analysed corpora, but the much higher volume of text will enable e.g. quantitative linguistic studies.

REFERENCES

Cinková, S., Toman J., Hajič J., Čermáková K., Klimeš V., Mladová L., Šindlerová J., Tomšů K. & Žabokrtský, Z. (2009). Tectogrammatical Annotation of the Wall Street Journal. *Prague Bulletin of Mathematical Linguistics*, 85-104.

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

Haverinen, K., Ginter, F., Laippala, V., Viljanen, T. & Salakoski, T. (2009). Dependency Annotation of Wikipedia: First Steps towards a Finnish Treebank. In Marco Passarotti

, Adam Przepiórkowski

, Savina Raynaud

and Frank Van Eynde

(Eds), *Proceedings of The Eighth International Workshop on Treebanks and Linguistic Theories (TLT8*) (pp. 95-107). Milano: EDUCatt

.

Jäppinen, H., Lehtola A. & Valkonen K. (1986). Functional structures for parsing dependency constraints. In *Proceedings of the 11th conference on Computational linguistics*. Association for Computational Linguistics (pp. 461-463). Bonn: Institut for angewandte Kommunikations- und Sprachforschung e.V.

Karlsson. F., Voutilainen, A., Heikkilä J. & Anttila A. (1995). *Constraint Grammar: A Language-Independent Framework for Parsing Unrestricted Text*. Berlin / New York: Mouton de Gruyter.

King, T., Crouch, R., Rietzler, S., Dalrymple, M. & Kaplan, R. M. (2003). The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*. Budapest: ACL.

Marcus, M., Santorini B. & Marcinkiewicz M. (2004). Building a large annotated corpus of English: the Penn Treebank. In G. Sampson & D. McCarthy (Eds.), *Corpus Linguistics: Readings in a Widening Discipline.* New York: Continuum.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Tapanainen, P. & Järvinen T. (1997). A non-projective dependency parser. In *Proceedings of the fifth conference on Applied natural language processing.* Washington, DC: ACL.

Voutilainen, A. (2003) Part-of-Speech Tagging. In Ruslan Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp 219-232). Oxford and New York: Oxford University Press.