Dependency Annotation of Wikipedia: First Steps Towards a Finnish Treebank

Katri Haverinen,[1,3] Filip Ginter,[1] Veronika Laippala,[2]
Timo Viljanen,[1] Tapio Salakoski[1,3]

[1]Department of Information Technology,
[2]Department of French studies
[3]Turku Centre for Computer Science (TUCS)
20014 University of Turku, Finland
first.last@utu.fi

Abstract

In this work, we present the first results obtained during the annotation of a general Finnish treebank in the Stanford Dependency scheme. We find that the scheme is a suitable syntax representation for Finnish, with only minor modifications needed. The treebank is based on text from the Finnish Wikipedia, ensuring its free distribution and broad topical variance. To assess the suitability of Wikipedia text as the basis of a treebank, we analyze its grammaticality and find the quality of the language surprisingly high, with 97.2% of the sentences judged as grammatical. The treebank currently consists of 60 fully annotated articles and is freely available.

1 Introduction

Treebanks are among the most crucial resources for the development of natural language processing (NLP) methods. There exist a number of national treebanks for a variety of languages, including widely used and studied ones, such as English, as well as languages spoken by comparatively smaller populations, for example Slovene. For Finnish, no such treebank currently exists, considerably restricting the possibilities for NLP research for this language. To address this obvious deficiency, we have commenced an effort to develop the first Finnish language treebank and, in this paper, present the first results of this project.

The source of text for the treebank is the Finnish Wikipedia. One of its major advantages is that it is released under a free license, enabling the distribution of the resulting treebank at no cost and with no copyright issues. Apart from offering a great topical variety, the text is written collaboratively by a number of authors and thus also reflects a number of different personal writing styles. Since there is little

prior work on Wikipedia-based treebanking, we assess the grammaticality of the language and thus, to some extent, its suitability for a source of treebank text.

The annotation scheme of the treebank is the well-known Stanford Dependency (SD) scheme which was designed specifically for NLP applications [1, 10]. The Finnish treebank is the first general language corpus annotated natively in the SD scheme. Since the scheme was originally designed for English, we discuss its applicability to Finnish as part of the results presented in this paper. In particular, we show that only minor modifications to the scheme are necessary. The choice of the scheme follows a recent substantial interest in the application of dependency schemes in general and the numerous successful applications of the SD scheme specifically [8, 10, 12].

Among the most important application areas for treebanks is the induction and evaluation of statistical parsers. For instance, a number of national treebanks for diverse languages such as Catalan, English, and Japanese have been used in the recent CoNLL'09 shared task [2] to develop and evaluate multilingual statistical parsers, thus greatly benefiting the NLP research for these languages. Indeed, one of the primary motivations for this work is to provide a similar opportunity for Finnish NLP research. This motivation has affected both the choice of the scheme and the target size of the corpus, as will be discussed later.

2 Related work

As stated earlier, there is no publicly available treebank of general Finnish. The only treebank we are aware of is that of Haverinen et al. [4] who have applied the SD scheme to Finnish intensive care nursing narratives, producing a treebank of 1019 sentences. This treebank, however, is not publicly available due to patient privacy issues.

Also other NLP resources for Finnish are scarce. The only broad-coverage full syntactic parser for Finnish is the closed source commercial parser Connexor Syntax.[1] Other NLP tools, particularly targeted at morphological analysis, include FinTWOL and FinCG,[2] a morphological analyzer and a Constraint Grammar parser which resolves morphological ambiguity [5, 6], both commercial products. In addition, a rule-based parser has been developed by Laippala et al. [7], particularly targeting the language used in nursing narratives in a Finnish intensive care unit. This parser is, however, restricted to the very specific vocabulary and syntax typical for this domain.

Apart from the nursing narrative corpus of Haverinen et al., there is a second treebank with SD as its native annotation scheme, BioInfer [13]. It is an English-language corpus of 1100 sentences from research article abstracts focusing on protein-protein interactions. In addition to these two corpora, any treebank

[1] `http://www.connexor.eu`
[2] `http://www.lingsoft.fi`

annotated in the Penn Treebank [9] scheme can be automatically converted to the SD scheme using the method and tools[3] of de Marneffe and Manning [10].

3 Adaptation of the SD scheme to Finnish

In this section, we introduce our modifications to the Stanford Dependency scheme. Sections 3.2 and 3.3 discuss Finnish-specific adjustments, while Sections 3.4 through 3.7 consider more general modifications. Due to space limitations, the original SD scheme will only be discussed briefly, and the reader is referred to the work of de Marneffe and Manning [1] for a thorough description.

3.1 The Stanford Dependency scheme

In the SD scheme, the syntactic structure of a sentence is represented as a directed graph of labelled dependencies. The latest scheme version [1] defines 55 hierarchically arranged dependency types, capturing both syntactic and semantic relations. There are four different representation variants, in which different sets of dependencies are present. In the basic variant, used in the current annotation, the analyses are trees and generally include only syntactic dependencies. Other variants define a number of additional, semantically motivated dependency types that are present in addition to the basic syntactic dependencies. These variants thus result in non-tree structures that may even contain directed cycles.

The scheme is designed to be application-oriented and has indeed proved its usefulness in a number of NLP methods (for an extensive list, see the review by de Marneffe and Manning [10]). These successful applications have also contributed to our decision of using the scheme in this work, as has the encouraging observation that the SD scheme would seem to be suitable at least for clinical Finnish, as reported by Haverinen et al. [4].

Haverinen et al. adapted the SD scheme to clinical Finnish by introducing several new dependency types that address the most common Finnish syntactic structures that the SD scheme could not naturally represent: inflected nominal modifiers, adpositional phrases, and certain passive structures (for details, see [4]). These modifications apply with no further changes also to general Finnish, and, in the following, we discuss our additional adaptations of the scheme.

3.2 Genitive objects

In Finnish, a noun with a verb counterpart or a nominalization of a verb can have an object, called the *genitive object*. This resembles the English phenomenon where a gerundial noun takes an object in front of it, as in *ship building*, except that the genitive case is not used in the English structure. In English, nominal pre-modifiers

[3] `http://nlp.stanford.edu/software/lex-parser.shtml`

such as the above are considered syntactic compounds and are marked *nn* in the SD scheme.

On the surface, genitive objects are identical to possessive modifiers, both being nominal pre-modifiers in the genitive case. There is, however, a clear semantic difference between these two. For instance, the possessive interpretation of *laivan rakentaminen* (*ship+genitive building*) would mean that the ship itself is doing the building, whereas the genitive object interpretation would mean that the ship is being built. In order to maintain this semantic distinction, it is necessary to establish a new dependency type, *gobj*, for genitive objects.

3.3 Finnish copulas

The SD scheme reserves a special treatment for copula structures: the predicative of a copular clause is the head and the copular verb its dependent. In all other cases, the finite verb acts as the head. This is motivated from a multilingual point of view, as not all languages have an overt copular verb. Further, particularly in telegraphic style, the copular verb can often be omitted even in those languages that do. This treatment of copula structures, however, requires an exact definition of the class of copular verbs and predicatives.

The SD scheme uses a list of English copular verbs defined in the Penn Treebank, including, among others, *to be*, *to resemble* and *to become*. According to Finnish Grammar [3, §891], the only Finnish copular verb is *olla* (*to be*), and all clauses with *olla* as the main verb can be classified as copular. This includes clauses where the predicative is inflected in a local case, such as *Paketti on Oulusta* (*The_package is from_Oulu*). However, if a structure such as this one is accepted as copular, a sentence with several possible predicatives, such as *Paketti on Oulusta ystävältäni* (*The_package is from_Oulu from_my_friend*) can easily be formed. Such a structure has no obvious dependency representation in the SD scheme, since the clause would have two head words. Another problem related to the predicative cases is that of distinguishing the copular verb *olla* (*to be*) and other, non-copular verbs that take as their argument a noun inflected in the same case as the argument of the verb *olla*. Consider, for example, *olla laulajana* (*to_be singer+essive*), *toimia laulajana* (*to_act as_singer+essive*) and *työskennellä laulajana* (*to_work as_singer+essive*). All three examples have the same surface syntactic structure, yet for instance the third example is certainly not a case of copula.

To avoid the class of copulas becoming unnecessarily broad, and syntactically and semantically diverse, we only allow nominative and partitive cases for noun and adjective predicatives, which permits us to restrict copular structures to those that include the only Finnish copular verb, *olla*. In addition to nouns and adjectives, for instance adverbs and even full clauses can act as predicatives. Our solution, including our use of the separate copula subject type, *nsubj-cop*, is similar to that in the clinical treebank of Haverinen et al., although some of the most problematic cases do not occur in the clinical language. For an illustration of our analysis of Finnish copula structures, see Figure 1.

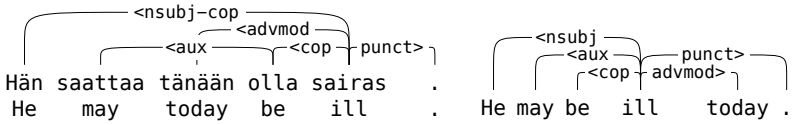

Figure 1: Finnish copula structures (left) as compared to those of English (right). Note that the copula acts as the head for the possible auxiliary which can sometimes cause non-projective structures. Also note the use of the *nsubj-cop* dependency type.

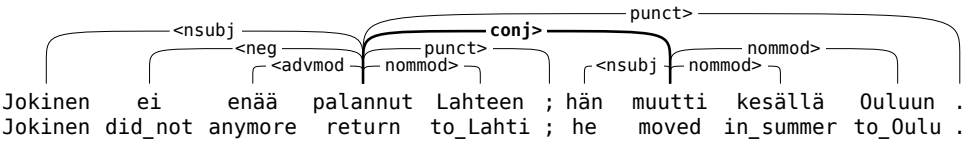

Figure 2: Implicit clausal coordination. The example sentence could be translated as "Jokinen did not return to Lahti anymore; he moved to Oulu in the summer."

3.4 Independent clause coordination

Independent clauses can be coordinated without a conjunction, as in *Lapset pyöräilivät kouluun; aikuiset ajoivat töihin* (*The children cycled to school; the adults drove to work*). The SD scheme analyzes such implicit coordination as parataxis and defines the corresponding dependency type. We, however find these structures functionally and semantically similar to explicit coordinations and thus also annotate them similarly. This is particularly natural in the SD scheme which analyzes conjunctions as mere dependents of the first coordinated element, making implicit and explicit coordinations differ only in the presence or absence of this single dependent. The *parataxis* type is then reserved for other types of parataxis such as reporting clauses. In this respect the scheme also diverts from that used by Haverinen et al., who defined a separate dependency type, *sdep*, for implicit clause coordination. Our analysis is illustrated in Figure 2.

3.5 Infinite clausal complements

The original SD scheme does not distinguish between finite and infinite clausal complements, but uses the type *ccomp* for both. For instance, in the structures *Sanoin, että pallo katosi* (*I_said that the_ball disappeared*) and *Estin palloa katoamasta* (*I_prevented the_ball from_disappearing*), the complements *että pallo katosi* (*that the_ball disappeared*) and *palloa katoamasta* (*the_ball from_disappearing*) would both be analyzed as *ccomp* in the original SD scheme. The *iccomp* dependency type enables the distinction of these two structures, which would otherwise not be possible without morphological information that, currently, is not present in the treebank.

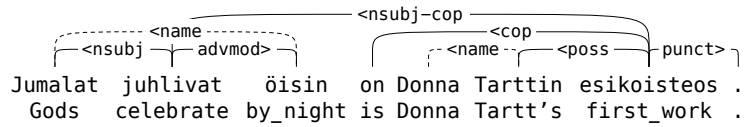

Figure 3: *Jumalat juhlivat öisin* (*Gods celebrate by night*) is a named entity with an inner syntactic structure and is thus given a full syntactic analysis, including the correct head word. *Donna Tarttin* is only marked as a multi-word unit with no further analysis. The technical dependency *name* is used to delimit named entity boundaries.

3.6 Named entities

Multi-word named entities, such as names of people, cities, books, and movies, are frequent in general language. These elements are problematic in a number of ways: often, but not always, they lack an obvious inner syntactic structure, despite consisting of several words, as for example *Carl Gustaf Emil Mannerheim*, or they may also be in another language, like *Västra Finnholmen*. An example of a name that does have a complex inner structure and is in Finnish would be the name of the book *Taistelu sosiaaliturvasta — ammattiyhdistysväen toiminta sosiaaliturvan puolesta 1957–1963* (in English *The battle for social security — trade union members' actions for social security 1957–1963*).

All multi-word names are annotated as single units whose rightmost word acts as the head in the dependency tree. In addition, Finnish names that do have an inner syntactic structure are given a full dependency annotation and their correct head word is identified (see Figure 3 for an illustration). This approach thus leaves open two options for treating Finnish named entities with inner structure. One possibility is to discard the annotation of the inner structure and consequently treat the named entities as single units. The other alternative is to preserve these entities as subtrees in the syntactic structure. The choice will likely be application-dependent.

3.7 Gapping and fragments

Gapping, a form of ellipsis where a governing element is omitted to avoid repetition while its dependents are not, poses an annotation problem. For instance, in *minä söin jäätelöä ja sinä salaattia* (*I ate ice cream and you salad*), the elided verb is necessary to construct a tree that correctly reflects the meaning of the sentence. A similar case is that of fragments, such as *Presidentti Kiinaan* (*The President to China*), where the head word of the clause is absent.

In order to be able to construct an analysis for such cases, we insert a *null* token into the sentences to represent the missing head word. In the case of gapping, where the antecedent of the elided element is present earlier in the sentence, we further include a semantic dependency, *ellipsis*, to relate the antecedent and the *null* token. In the case of fragments, no antecedent is present in the sentence and consequently

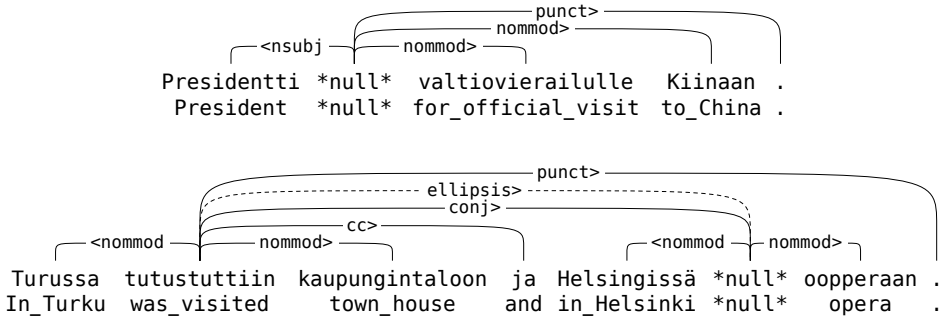

Figure 4: Null tokens in the case of fragments (top) and gapping (bottom). Note the semantic dependency *ellipsis*. The fragment sentence could be translated as "The President for official visit to China" and the ellipsis sentence as "The town house was visited in Turku and the opera in Helsinki."

the *ellipsis* dependency is not used. Figure 4 illustrates the usage of *null* tokens.

Note that the *null* tokens are only used to stand for missing governors. Consequently, other elements that do not generally act as governors in the SD scheme, such as missing copula verbs and auxiliaries, are not represented by *null* tokens, neither are other forms of ellipsis.

4 Construction of the Treebank

In this section, we describe the ongoing work on the Finnish dependency treebank itself.

4.1 Treebank text

In constructing the treebank, we use randomly selected articles from the Finnish Wikipedia. All articles that do not exceed 75 sentences are annotated in their entirety, excluding parts that do not have enough syntactic structure to annotate, such as bulleted lists of single words, section headings and figure captions. Longer articles are truncated at 75 sentences, to keep the treebank from becoming biased towards a single article topic.

Currently, we have completed the annotation of 60 articles, comprising 711 sentences and 10217 tokens, of which 8801 are non-punctuation. Thus, on average, one article is 11 sentences long and one sentence contains 14 tokens. The length of articles varies substantially — the longest article in the currently annotated part is 61 sentences long, while the shortest contains only one sentence. Out of all sentences in the treebank, 27 (3.8%) are non-projective. For comparison, Haverinen et al. report the proportion of non-projective sentences in the clinical treebank to be 2.9%. The currently existing annotation, subject to further changes, is available at `http://bionlp.utu.fi/fintreebank.html` to illustrate

the annotation scheme.

4.2 The Annotation process

In our annotation work, we use a custom annotation tool, which will be made publicly available together with the treebank. It includes the basic abilities necessary for dependency annotation, along with search abilities and the possibility to mark a dependency for later discussion, or label a sentence as dubious or ungrammatical.

We have started the annotation process by first annotating 562 sentences (47 articles) in trial annotations. Each sentence was first annotated by one annotator and the annotation was then jointly inspected by the whole group. Authoritative decisions were made for all problematic cases found at this stage, and the already existing annotation was modified as necessary to ensure its consistency.

After the trial annotations, full double annotation has been started. That is, each sentence is first independently annotated by two annotators and all differences are then jointly resolved. The decisions made at this double annotation stage lean on the authoritative decisions made after the trial annotations. The current number of double annotated sentences is 149 (13 articles). Due to the currently small number of these sentences, we do not report an inter-annotator agreement at this stage, as this figure would not be representative. Inter-annotator agreement in the double annotation will be measured regularly throughout the annotation process to estimate the annotation quality and will be reported with the final release of the corpus.

5 Characteristics of Wikipedia text

The text in Wikipedia articles is sometimes thought to be of poor quality with respect to grammaticality. To determine some properties of the Wikipedia language, we have conducted a small-scale analysis of the currently annotated sample, estimating the proportion of spelling and grammar errors.

We assess the amount of spelling errors in the text by manually inspecting all words that FinTWOL,[4] a broad-coverage morphological analyzer, failed to recognize. Of the 1034 (10.1% of all tokens) unrecognized tokens, only 6 (0.6‰) were obvious misspellings, the remaining being most commonly names, foreign words, numerical expressions, untypical punctuation symbols, abbreviations, etc.

To estimate the level of ungrammaticality in the Wikipedia text, each sentence was assessed independently by three native speaker annotators, and marked *grammatical*, *questionable* or *ungrammatical*. All sentences not judged grammatical by at least two of the three annotators were further manually analyzed to determine the type of error they contained. The results of this manual analysis are given in Table 1. The vast majority of sentences, 691 out of 711 (97.2%), were judged grammatical by at least two annotators; 627 (88.2%) were judged grammatical unanimously. Further, 18 sentences (2.5%) were judged questionable and

[4] `http://www.lingsoft.fi`

Mistake type	Frequency
Fragment	6
Relative clause error	4
Compound error	3
Translation error, anglicism or colloquial	6
Inflection error	2
Coordination error	2
Total	23

Table 1: Results of the manual analysis of grammar errors. Note that the total number of errors is greater than the total number of ungrammatical and questionable sentences, as some sentences had more than one error in them.

2 (0.3%) ungrammatical. Out of the 20 sentences not judged grammatical, only one was downright incomprehensible. Fragments are among the most common cases judged questionable or ungrammatical, as are translation errors, anglicisms and colloquial language.

In general, many sentences judged as questionable were colloquial rather than strictly erroneous. Examples of such colloquial structures, which would in some contexts be judged ungrammatical, can be a sizeable asset for example when building a parser targeting text produced by non-professional writers. To conclude, we find the overall quality of the Wikipedia text, in terms of grammaticality and correct spelling, clearly acceptable.

6 Conclusions and future work

In this paper, we have presented first results of an ongoing effort to build a treebank of the Finnish language. First, we demonstrate that the Stanford Dependency scheme is applicable to general Finnish with only minor modifications. Many of these modifications have previously been introduced by Haverinen et al. [4] who applied the SD scheme to Finnish nursing narratives. Second, we assess the grammaticality of the Finnish Wikipedia language and find it, maybe somewhat surprisingly, clearly acceptable. In addition to the obvious benefit that Wikipedia text is freely available under an open license, it may also be an asset for a number of real-world applications that the language found in the articles can be colloquial and is not necessarily produced by professional writers. Currently, the treebank consists of 60 fully annotated articles, comprising of 711 sentences. The annotation is available at `http://bionlp.utu.fi/fintreebank.html`.

The primary goal of the project is to create a freely available treebank large enough for the induction of a broad-coverage statistical parser as well as the development of natural language processing methods in general. The first and most important future work direction is thus naturally to increase the size of the corpus.

Currently, we aim at annotating roughly 10,000 sentences, that is, about 140,000 tokens, a treebank size shown to be sufficient to induce an accurate statistical parser for a number of languages [11]. The performance and learning curve of the induced parser and other NLP methods that use the treebank will help to determine its final size.

A second, more long-term direction is to further enhance the annotation of the treebank by providing a layer of more detailed semantic analysis, for example using an SD scheme variant that also includes semantically oriented dependency types. In this layer, it would also be possible to deepen the annotation of elliptic structures by marking also omission of non-head elements. This will require further modifications to the SD scheme which does not prescribe any treatment of ellipsis. Thirdly, the possibility to provide morphological and POS information for the treebank using an existing analyzer for Finnish will be investigated.

Acknowledgements

We would like to thank Lingsoft Ltd. for making FinTWOL available to us. This work has been supported by the Academy of Finland and Turun Yliopistosäätiö.

References

[1] Marie-Catherine de Marneffe and Christopher Manning. Stanford typed dependencies manual. Technical report, Stanford University, September 2008.

[2] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL 2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL'09*, pages 1–18, 2009.

[3] Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho. *Iso suomen kielioppi / Grammar of Finnish*. Suomalaisen kirjallisuuden seura, 2004.

[4] Katri Haverinen, Filip Ginter, Veronika Laippala, and Tapio Salakoski. Parsing clinical Finnish: Experiments with rule-based and statistical dependency parsers. In *Proceedings of NODALIDA'09, Odense, Denmark*, 2009.

[5] Fred Karlsson. Constraint Grammar as a framework for parsing unrestricted text. In *Proceedings of COLING'90*, pages 168–173, 1990.

[6] Kimmo Koskenniemi. Two-level model for morphological analysis. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 683–685, 1983.

[7] Veronika Laippala, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. Towards automatic processing of clinical Finnish: A sublanguage analysis and a rule-based parser. *International Journal of Medical Informatics, Special Issue on Mining of Clinical and Biomedical Text and Data*, 2009. In press, available in online version only.

[8] Dekang Lin. A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4(2):97–114, 1998.

[9] Mitchell Marcus, Mary Ann Marcinkiwicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2), 1993.

[10] Marie-Catherine de Marneffe and Christopher Manning. Stanford typed dependencies representation. In *Proceedings of COLING'08, Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, 2008.

[11] Joakim Nivre. Deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553, 2008.

[12] Joakim Nivre. Sorting out dependency parsing. In *Proceedings of GoTAL'08*, pages 16–27, 2008.

[13] Sampo Pyysalo, Filip Ginter, Veronika Laippala, Katri Haverinen, Juho Heimonen, and Tapio Salakoski. On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA. In *Proceedings of BioNLP'07*, pages 25–32, 2007.