

Defining the Annotation Scheme of a Treebank: The End-Use Perspective

Kristiina Muhonen & Tanja Purtonen

University of Helsinki
Department of Modern Languages
FIN-CLARIN
{kristiina.muhonen, tanja.purtonen}@helsinki.fi

Abstract

In this paper we discuss the importance of the end-use perspective when designing the annotation scheme of a treebank. Treebank usage reflects the value of the treebank, and should be considered when evaluating them. We discuss the issues rising from defining an annotation scheme from the point of view of FinnTreeBank, a dependency treebank for Finnish. These issues include deciding where to draw the annotation scheme from, what levels to include in it, and how to fit the levels together. To maximize the utility of the treebank, all levels of annotation should conform to the needs set by the treebank’s users.

Keywords: treebank, annotation scheme, dependency syntax

1. Introduction

Roughly put, completed treebanks have no intrinsic value. The value of a treebank is determined by its usage. The usability of a completed treebank should mainly be considered from the perspective of its end-users, be they researchers or an NLP application. Hence, the real value of a treebank is instrumental, it is the solely the end-users who determine whether the treebank is useful.

In this paper we discuss the importance of choosing a well-suited annotation scheme for a treebank so that it supports the end-use. The paper describes work-in-progress, namely the issues that need to be addressed in the early phases of developing the annotation scheme. We describe the process from the point of view of a Finnish dependency treebank, FinnTreeBank¹ (Voutilainen and Lindén, 2011). Assessing the usage and usability of a treebank is so far an unresolved problem. At this point, during the building phase of FinnTreeBank, we approach the problem only from the annotation scheme’s perspective and discuss what demands end-use sets on the scheme.

We first discuss treebank evaluation in general and then report some usage statistics of existing treebanks. In Section 3. we discuss the process of defining the annotation scheme and the interaction between the different levels of annotation. In Section 4. we explore the different sources for the scheme and show what effect the choice has for the use of the treebank.

2. Usage-Based Treebank Evaluation

Usage reflects the value of a treebank and should be considered in treebank evaluation. By keeping the end-use perspective in mind already at the very start of a treebank effort, we try to avoid building a low-utility resource. In the following section we sketch how treebanks could be

evaluated from a usage point of view. We start by discussing different annotation decisions and conclude with briefly scrutinizing the usage of existing treebanks.

2.1. Evaluating Annotation Decisions

Traditionally, treebank evaluation focuses on coverage and accuracy. (For methods for qualitative comparison of treebanks, see e.g. Kübler et al. (2008).) In addition to the traditional treebank evaluation methods, the creators of treebanks can acquire important knowledge on the usage and usability of treebanks by collecting and analyzing user statistics as well.

From the user’s perspective, the choice of analysis makes a big difference. However, these choices cannot be measured with traditional LAS scores. For example, there are two different ongoing Finnish dependency treebank projects, FinnTreeBank and the Turku Dependency Treebank (TDT). Both projects have different kinds of ways for specifying postpositions. This means that the word *kuluttua* (*passed/in. . . time*) in the example below is analyzed in two different ways.

- (1) tunnin kuluttua
hour-GEN passed-PAST-PRC-PASS/in. . . time
TDT: *when an hour has passed*
FinnTreeBank: *in an hour*

The word *kuluttua* (*passed/in. . . time*) is analyzed in two different ways: as a verb and as a postposition. Nonetheless, when the modeling is once specified, the creators of the treebanks can publish error rates and LAS-scores without clarifying in which way the “correct” analysis is the right one.

Regardless of which reading of *kuluttua* is the “correct” one, the user determines whether the analysis is useful for her purposes. For example, a researcher querying the corpus for adpositional phrases benefits more from

¹<http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/>

kuluttua (passed/in...time) being analyzed as a postposition, where as a researcher studying participles prefers the verbal reading. The annotation scheme of FinnTreeBank marks *kuluttua* in Example 1 as a postposition. This choice, like the whole annotation scheme, is based on a wide-coverage descriptive grammar of Finnish (Hakulinen et al., 2004a). Therefore we assume that the postpositional reading is adequate for research purposes and that it follows the consensus within the research community.

Independent of the treebank users, also the complexity of the annotation scheme and inter-annotator agreement influence the feasibility of the treebank, and therefore also inherently the reception of the treebank. These factors have a great impact on the usability of the treebank, and can be measured in traditional ways like LAS-scores.

The annotation scheme of FinnTreeBank has recently been tested in Voutilainen and Purtonen (2011). The authors assess the complexity of the annotation scheme by double-blind tests. They hypothesize that inter-annotator agreement suffers if the annotation scheme is too complex, but conclude that based on the tests, the complexity of scheme used in FinnTreeBank does not pose this problem.

2.2. User Statistics

The usage of a treebank can e.g. be seen from the amount of NLP projects relying on the treebank and the number of researchers using the treebank. To get an overview on the user statistics of ongoing treebank projects, we asked the developers to distribute user statistics via the Corpora mailing list². Since user statistics are gathered in several different ways and cannot be directly compared, the results displayed in Table 1 are merely illustrative.

Comparing the raw download counts or the number of licenses/registered users of different treebanks is difficult. The distribution policies differ from open-source to paid licenses, and gathering user statistics varies accordingly. For instance, the Copenhagen Dependency Treebank (Kromann, 2003) is open-source and does not gather any download statistics. The Icelandic Parsed Historical Corpus (Ingason et al., 2011) is also distributed freely and reports as many as 640 downloads of all five versions of the treebank.

Counting the exact number of users of open-source treebanks is problematic because users are free to distribute the data without reporting it. Also, download counts are distorted by issues like users downloading the data several times. For instance, the freely available Turku Dependency Treebank (TDT) (Haverinen et al., 2010) has 35 unique registered users, but the raw download counts from server logs are higher.

The Quranic Arabic Corpus is also freely available but only for online viewing and search. The project reports more than 2500 daily visitors mostly using the POS and morphologically tagged corpus.

The Prague Dependency Treebank (Hajič, 1998) is distributed under a paid license or through the Linguistic Data Consortium (LDC). Since the licenses can be signed

by institutions, estimating the actual user number is hard. This problem applies to open-source treebanks as well.

Another way of assessing the usage of a treebank is to study citation statistics. A quick-and-dirty method for doing this is to search Google Scholar for articles which refer to the treebank. The results of these queries are also reported in Table 1.

In addition to it being hard to know the exact number of treebank users or research relating to them, the difference in the size of both the language and the treebank, and age of the treebank also make quantitative comparison of treebanks difficult. However, since the value of a treebank is instrumental, also usage should be considered when evaluating a treebank.

We will now move away from quantitative comparison and proceed to discuss how the annotation scheme of a treebank adds to the value of a treebank and makes it more usable.

3. Choosing the Annotation Scheme

Treebank usage mirrors the success of the annotation scheme: If the scheme is not chosen according to the needs set by the end-users, the treebank will not be supported by the research community. User-driven treebank design requires making decisions about organizing information on different levels and determining how the levels interact with each other in a harmonious way.

3.1. Annotation Levels

When a treebank has two or more levels of annotation, treebankers must decide on which level they represent which information. Multiple levels make it possible for each level to take its own perspective independently without being influenced by other levels. However, when aiming at building a treebank for research purposes, the different levels should not be based on contradicting theories. For example, OMorFi, a morphological analyzer/generator for Finnish, is based on theory where verb derivatives are seen as verbs. However, in many syntactic theories, only nominal phrases can be seen e.g. as a subject (Hakulinen et al., 2004b).

In FinnTreeBank, the main focus is on the syntactic level and the morphological annotation scheme is designed from a syntactic point of view. In the annotation scheme, this can be seen e.g. in the treatment of derived nouns:

(2)	Postimerkkien keräily	on	kivaa.
	stamps	collecting	fun
	OBJ	SUBJ	PRED SCOMP
	N	N+V_DER	V ADJ
	<i>Collecting stamps is fun.</i>		

As can be seen in Example 2, on the syntactic level, only a nominal phrase (be it a noun or e.g. an infinitive structure) can be seen as a subject. On the morphological level, this syntactic annotation principle leads to a solution where the part-of-speech tag of *keräily* (*collecting*) is noun (N). However, the information about the derivation is visible on the morphological level as well: V_DER.

²<http://www.hit.uib.no/corpora/>

TREEBANK	USAGE ¹	NOTES	GOOGLE SCHOLAR**
Bulgarian Treebank <i>www.bultreebank.org/</i>	64 9	CoNLL 2006 HPSG	312
Prague Dependency TB 1+2 (LDC)	365	v 1.0 & v 2.0	1200
Icelandic Parsed Historical Corpus	640 *	5 versions	7
Turku Dependency TB	35*	unique registered users	2
Copenhagen Dependency TB	NA	no statistics	16
Quranic Arabic Corpus <i>http://corpus.quran.com/</i>	2500*	online users, visitors/day	19
Penn Treebank 2 (LDC)	~150		8850
Penn Treebank 3 (LDC)	~250		
Penn Chinese TB 7.0 (LDC)	~600	all versions	609
Penn Arabic TB 1 (LDC)	~300	v 2.0, 3.0 & 4.1	221
Penn Arabic TB 2 (LDC)	~100	v 2.0	
Penn Arabic TB 3 (LDC)	~300	v 1.0, 2.0 & 3.2	
Penn Arabic TB 4 (LDC)	~100	v 1.0	
* = Free			
** 25 October 2011			
¹ Users/Downloads/Licenses			
LDC = Linguistic Data Consortium, <i>www ldc.upenn.edu/</i>			

Table 1: User Statistics of Different Treebanks

FinnTreeBank aims at a semantically informative annotation scheme. Thus, the NP-internal relations in *postimerkkien keräily* (*collecting stamps*) should be apparent. When the word *keräily* (*collecting*) is seen as a verb-to-noun derivation, like in Example 2, it is possible to mark *postimerkkien* (*stamps*) as its object. The word *postimerkkien* (*stamps*) is the object of *keräily* (*collecting*), which becomes apparent if the expression is transformed to a finite clause *kerätä postimerkkejä* (*to collect stamps*).

Even though nouns are not generally seen as possible heads of objects, the information about the noun's verb-based origin on the morphological level supports the semantically more precise interpretation where *postimerkkien* is seen as the object of the verb-based noun *keräily*. This requires adding the derivation information in the morphological analysis of *keräily* (*collecting*), V_DER.

3.2. Annotation Level Interaction

We use a rule-based formalism, Constraint Grammar (CG) (Karlsson et al., 1995), for creating FinnTreeBank. It enables the interaction between the annotation levels. By this we mean that based on the syntactic analysis we can add information to the morphological level. This can be useful e.g. in the sentences in which the subject is an independent adjective-like word that refers to a person:

- (3) Onnelliset elävät pisimpään.
happy live longest
Happy people live the longest.

In Finnish, every adjective can be used as a subject, like in Example (3). When designing the annotation scheme for Finnish, we need to decide, whether an adjective-like word referring to a physical object e.g. to a person should be analyzed as a noun or an adjective.

In FinnTreeBank, we approach this from the syntax's point of view, and on the syntactic level, only nominal phrases can function as a subject. However, when we have information about an adjective-like word functioning as a subject, we can add information to the morphological level after the syntactic analysis is completed. Hence, the word *onnelliset* (*happy*) is marked with the tag A_NP, which indicates that the adjective is used as an independent NP. In practice we accomplish this with CG, basing the context conditions of the rules on both syntax and morphology. With this approach we want to ensure that the scheme is as informative as possible and that the different annotation levels do not conflict with each other.

In the end, regardless of how this adjective-as-a-subject problem is solved, the most important thing from the user's perspective is to know, how these kinds of problems are solved in the treebank and from which level she can find the information needed. Hence, documenting the annotation decisions is crucially important, especially in the less obvious cases. For this purpose, FinnTreeBank provides a user manual, where the annotation scheme is defined (Voutilainen et al., 2011).

4. Where Does the Scheme Come from?

Even though FinnTreeBank is created in the dependency syntactic framework, it is not obvious how to model the dependency relations and grammatical functions in practice. On the syntactic level e.g. the dependency syntactic function palette and linking of the dependency relations need to be specified.

To ensure that the chosen linguistic modeling is relevant to the purposes of the treebank, it is worth considering from what source it is drawn. If the needs of the users are known, defining the annotation scheme is straightforward.

This is the case e.g. if one builds a treebank for NLP purposes and knows the application and its needs in advance. The application imposes constraints on the scheme and the scheme can be fitted for the application. When building a treebank for a vaguer purpose, e.g. research, designing the annotation scheme is not a self-evident task. We have to make assumptions on the users' needs and base the annotation scheme on them when the solution is not defined precisely enough in the descriptive grammar (Hakulinen et al., 2004a). In the building phase, there is no easy way to test these postulations, so they need to be formulated carefully.

Among the first decisions that need to be made when building a treebank from scratch is where to draw the annotation scheme from. There are at least four main sources for the annotation scheme: (1) another treebank, (2) a grammar, (3) own research, and (4) consulting the users. These ways can be, and usually are, combined.

Source 1: Another Treebank

When creating the first treebank for a language, the annotation scheme can be drawn from a treebank for some other language. The advantage of this approach is compatibility and effortlessness. TDT employs the Stanford Dependency scheme (de Marneffe and Manning, 2008a). They found that the scheme originally developed for English can be used for Finnish with only minor modifications (Haverinen et al., 2010). In fact, as Buch-Kromann (2010) notes, the future challenges in treebanking lie in conversion of annotation schemes, not in finding a fit-all scheme.

The disadvantage of using a scheme made for another language is that when the scheme is fitted to the target language, some structures not present in the source language can be ignored consciously or unconsciously. The scheme needs to be flexible enough to allow incorporation of structures not present in the original language. Detecting the structures that do not fit the original scheme requires scrutinizing a large amount of text.

If a treebank is built for the purposes of language researchers, the best annotation scheme might not be the same as when a treebank's primary purpose is to support NLP. In de Marneffe and Manning (2008b) the authors give an example of trading off linguistic fidelity for representability. The widely used Stanford typed dependencies representation sometimes produces an analysis which is not attractive to a user conducting linguistic research. The authors give an example of the interaction between preposition collapsing and PP conjunction that requires predicate copying when using the Stanford dependencies. E.g. the sentence *Bill went over the river and right through the woods* is transformed into a sentence with VP coordination, and the verb *went* is copied: *Bill went over the river and went right through the woods* (de Marneffe and Manning, 2008b).

FinnTreeBank does not adopt the Stanford copying approach because we assume that a linguist doing research on elliptical coordination structures does not benefit from such a representation. Ideally the texts linguists base their research on should not be altered in any way. The primary

reason for using annotated corpora is to explore phenomena present in authentic language.

The VP-coordination example demonstrates that adjusting an existing annotation scheme to fit the needs of a new treebank can be problematic. If the linguistic notions of e.g. coordination or crossing branches differ, the scheme is not adoptable as such.

Source 2: A Grammar

The annotation scheme can be based on a descriptive grammar. E.g. in Finnish there is no dependency syntactic grammar among the up-to-date grammars. Using the newest extensive descriptive grammar of Finnish, Hakulinen et al. (2004a), as the base for the dependency syntactic representation of the treebank requires building the dependency relations from scratch.

Creating the dependency syntactic annotation scheme according to a non-normative grammar is more laborious than modifying an existing scheme. However, it ensures that the resulting treebank complies with the standard grammar of Finnish.

The most important advantage of using a descriptive grammar as a source of the annotation scheme is that the linguistic modeling is language-specific, and the nature of the language with its specialities are taken into account in the scheme. This approach is therefore especially suited for a treebank built for linguistic research, e.g. FinnTreeBank, because the treebank has a solid linguistic background and the grammar has already been validated by the user community.

Sources 3 & 4: Research and Consulting Users

The third way of defining the scheme is to develop it itself. As Buch-Kromann (2010) notes, building a theory-neutral treebank is problematic. The annotation scheme cannot only be based on the annotators' intuition; a meaningful annotation scheme is always based on some notion of linguistic theory. Even if the scheme is developed from scratch for the purposes of the treebank, it is always influenced by previous schemes or linguistic theories.

The fourth way of constructing the annotation scheme is consulting the future users of the treebank. This way the users' opinion can have an effect on the definition of the scheme. FinnTreeBank will provide annotated text first and foremost to the purposes of linguistic research. Thus, the scheme should not contradict with the current consensus of language research of Finnish. A pilot study on gathering the users' opinion on complex syntactic phenomena at the grammar definition phase is presented in (Muhonen and Purtonen, 2011).

5. Conclusion

In this paper we have discussed the importance of user-driven annotation scheme design. If the scheme used in the treebank is not suitable for the treebank's purpose, or if the users of the treebank find the annotation solutions unusable, the treebank has no value. Treebanks should be seen as a resource for the users, and they should be developed from the perspective the users' needs.

Even though a treebank would be built keeping the users in mind, it is hard to evaluate the success of the effort. User statistics do not tell everything about the use of treebanks. They are also not comparable, because treebanks are distributed under different licenses, are made for big and small languages, and are of varying ages. Hence, quantitative comparison of treebank use is virtually impossible.

In this article we have examined user-centered annotation scheme design. The treebank's end-use has an impact on where to draw the annotation scheme from and what information should be present on each level of annotation. We have demonstrated that CG enables interaction between the different annotation levels and concluded that the documenting the annotation decisions well is essential for user-friendly treebanking.

At the beginning of building a treebank, it is most important to define the annotation scheme based on educated assumptions on what the end-use of the treebank demands from the scheme. At a later phase, when the treebank is completed, and we can gather user feedback and user statistics, we can concretely evaluate whether our assumptions were true and figure out means to correct the false ones.

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 Atro Voutilainen and the three anonymous reviewers for their constructive comments.

6. References

Buch-Kromann, M. (2010). Open challenges in treebanking: some thoughts based on the Copenhagen Dependency Treebanks. In: *Workshop on Annotation and Exploitation of Parallel Corpora AEPC 2010*, Volume 10 of *NEALT Proceedings Series*, Tartu: Tartu University, pp. 1–13.

de Marneffe, M.C. and Manning, C.D. (2008a). *Stanford typed dependencies manual*. Stanford: Stanford University.

de Marneffe, M.C. and Manning, C.D. (2008b). The Stanford typed dependencies representation. In: *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1–8.

Hajič, J. (1998). Building a syntactically annotated corpus: The Prague dependency treebank. In: *Issues of valency and meaning*, pp. 106–132.

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

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. and Alho, I. (2004b). *Ison suomen kielioopin verkkoversio: määritelmät*. Suomalaisen Kirjallisuuden Seura. Retrieved from:

<http://kaino.kotus.fi/cgi-bin/visktermit/visktermit.cgi>. Access date: July 15, 2011

Haverinen, K., Viljanen, T., Laippala, V., Kohonen, S., Ginter, F. and Salakoski, T. (2010). Treebanking Finnish. In: Dickinson, M., Määrisep, K., and Passarotti, M. (Eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pp. 79–90.

Ingason, A.K., Wallenberg, J.C., Sigurðsson, E.F. and Rögnvaldsson, E. (2011). *Icelandic parsed historical corpus (IcePaHC)*. Retrieved from: http://www.linguist.is/icelandic_treebank. Access date: July 31, 2011

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. (Eds.). (1995). *Constraint Grammar: A Language-Independent System for Parsing Running Text IV* (Book Series: Natural Language Processing). Berlin and New York: Mouton de Gruyter.

Kromann, M. (2003). The Danish dependency treebank and the DTAG treebank tool. In: Nivre J, and Hinrichs, E. (Eds.), *2nd Workshop on Treebanks and Linguistic Theories (TLT)*, pp. 217–220.

Kübler, S., Maier, W., Rehbein, I. and Versley, Y. (2008). How to compare treebanks. In: *Proceedings of LREC 2008*, Volume 8.

Muhonen, K. and Purtonen, T. (2011). Creating a dependency syntactic treebank: Towards intuitive language modeling. In: Gerdes, K., Hajičová, E. and Wanner, L. (Eds.), *Proceedings of the International Conference on Dependency Linguistics*, pp. 155–164.

Voutilainen, A. and Lindén, K. (2011). Designing a dependency representation and grammar definition corpus for Finnish. In: Candel Mora, M.Á. and Carrió Pastor, M. (Eds.) *Proceedings of III Congreso Internacional de Lingüística de Corpus (CILC 2011)*, pp. 151–158.

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

Voutilainen, A., Purtonen, T., Leisko-Järvinen, S., Kumlander, M. and Muhonen, K. (2011). *Finnish grammar corpus and dependency syntax description*. Helsinki: University of Helsinki, Department of Modern Languages. Retrieved from: <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/>. Access date: July 31, 2011