Creating a Dependency Syntactic Treebank: Towards Intuitive Language Modeling

Kristiina Muhonen and **Tanja Purtonen**
Department of Modern Languages
University of Helsinki
kristiina.muhonen@helsinki.fi, tanja.purtonen@helsinki.fi

Abstract

In this paper we present a user-centered approach for defining the dependency syntactic specification for a treebank. We show that by collecting information on syntactic interpretations from the future users of the treebank, we can model so far dependency-syntactically undefined syntactic structures in a way that corresponds to the users' intuition. By consulting the users at the grammar definition phase we aim at better usability of the treebank in the future.

We focus on two complex syntactic phenomena: elliptical comparative clauses and participial NPs or NPs with a verb-derived noun as their head. We show how the phenomena can be interpreted in several ways and ask for the users' intuitive way of modeling them. The results aid in constructing the syntactic specification for the treebank.

1 Introduction

Building a treebank is an expensive effort consuming a lot of time and resources. To ensure the usability of the result, it is wise to ascertain that the chosen syntactic modeling responds to needs of its users. The Finnish CLARIN, FIN-CLARIN, project[1] provides language resources for researchers by creating a treebank and a dependency parser for unrestricted text. Because the main user groups of the Finnish treebank are presumably language researchers and students, it is necessary to ensure that the syntactic modeling used in the treebank accords with their linguistic intuition. In this paper we present a case study of improving the syntactic representation of the Finnish treebank on the basis of its user groups' judgment.

The FIN-CLARIN treebank project[2] is in a phase in which the first specification of the dependency syntactic representation and the first manually annotated FinnTreeBank are ready, and the morphological definition is in progress (Voutilainen and Lindén, 2011). The base for the first version of the treebank is a descriptive grammar of Finnish (Hakulinen et al., 2004a). The treebank consists of the grammar's example sentences[3]. The advantage of this approach is that already in the first version of the treebank every phenomenon described in the grammar must also be described in the dependency syntactic framework.

During the creation of the first treebank and the syntactic specification, the annotators encountered some phenomena in which it was hard to define the one and only best dependency syntactic representation. The problems in defining such phenomena are due to two reasons. Sometimes the descriptive grammar did not state only one specific representation for a phenomenon. In other cases the annotators reported that the traditional way of representing a phenomenon covered only the most typical cases but that the traditional representation seemed uninformative and unsuitable for covering the whole phenomenon.

In this paper we concentrate on two complex syntactic structures for which the wide-coverage descriptive grammar of Finnish (Hakulinen et al., 2004a) does not offer a complete solution: elliptical comparative clauses and NPs with either a participial construction or a verb-to-noun derivation. The two structures are only roughly defined in the first version of the treebank, and they need to be fully formulated in the second version. We

[1] http://www.ling.helsinki.fi/finclarin/

[2] http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/

[3] The online version of the grammar: http://kaino.kotus.fi/visk/etusivu.php

show that the dependency syntactic representation of undefined or complex structures can be better outlined when consulting the user groups of the treebank for their intuitive solution at the syntactic definition phase.

The user-centered approach guarantees that the syntactic representation complies with the majority's view which ensures maximizing the usability of the treebank. For this purpose we composed an e-query, in which we collected the answerers' intuitive interpretations of the two structures. Recording the user groups' intuitive solution complements, but does not replace the approximate syntactic representation already created in the project.

The first purpose of our experiment is to see how native speakers interpret elliptical comparative sentences, participial NPs with sentence-like structures and NPs with a verb-derived head. This sheds light on how the complex phenomena can be parsed in a natural way. The second aim is to estimate, is it beneficial to use an e-query at the syntactic specification phase. In this estimation we consider the number, the quality and the distribution of the answers. The third benefit of the test is to see whether there is a hidden consensus on the phenomena uncovered in the descriptive grammar and not yet described in the dependency syntactic framework. This, however, is not the main focus of our pilot study, but rather a side-product of the experiment.

2 Linguistic Background

In this section we outline the linguistic phenomena. We also show why the phenomena have alternative solutions.

2.1 Elliptical Comparative Sentences

The first phenomenon we concentrate on is the elliptical comparative structure. Finnish and English comparative structures are formed in a rather similar way. Typically a Finnish comparative structure contains:

- the comparative form of an adjective or an adverb formed with the comparative ending *-mpi*,
- the item being compared (subject of the main clause), and
- the subordinating conjunction *kuin*.

The next example shows a typical comparative structure:

(1) Ana on pidempi kuin Maria.
Ana is taller than Maria
Ana is taller than Maria.

In example (1) the target of the comparison is Maria and the item being compared is Ana. It is also possible that the target is not semantically equivalent with the item being compared, like in the following example:

(2) Ana on (nyt) pidempi kuin ennen.
Ana is (now) taller than before
Ana is now taller than before.

In this sentence, Ana is still the item being compared, but the comparative clause (*ennen/before*) is not comparable with the subject of the main clause (*Ana*), but with another word (*nyt/now*) in the previous clause. This equivalent word (*nyt/now*) is not necessarily even mentioned.

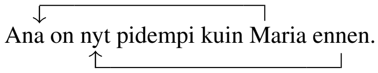
The diversity of comparative structures is a challenge for parsing: semantically oriented dependency parsing aims at an analysis in which the head is semantically, not only grammatically, considered the head. In our experiment, we investigate should sentences (1) and (2) be analyzed similarly with each other by marking e.g. the adjective, verb or the conjunction as the head. The other option is to link two equivalent words (e.g. *Ana–Maria, now–before*) with each other.

The comparative conjunction *kuin* can be followed by a whole, or an elliptical, sentence:

(3) Ana on nyt pidempi kuin Maria ennen.
Ana is now taller than Maria before
Ana is now taller than Maria before.

The comparative clause can be seen as a common structure of its own or as an elliptical clause. In principle, all cases where the comparative conjunction is not followed by a verb are elliptical clauses. In Finnish it is common to have a whole elliptical sentence after the comparative conjunction, like in example 3. Thus, the way of analyzing the comparative clause is significant; it can be analyzed as a structure of its own, or as an elliptical clause. In the tradition of dependency grammar, the subordinate clauses are linked to the main clause via the verb and all other head-dependent-relations stay inside the subordinating clause (Tesnière, 1980, p. 231). If the words following the comparative conjunction are seen as a clause, it is justifiable to have only one link from this clause to the main clause also in elliptical structures .

It is also possible to see the comparative as a conventional structure with a) no need to link the word following the conjunction to the main verb or b) no need to have only one link to the main clause. Thus the head-dependent relations can be seen e.g. in the following way (for the glossed sentence, see example (3)):



(4) Ana on nyt pidempi kuin Maria ennen.

In our experiment, we try to find out the most natural and informative way to describe different kinds of comparative structures. The main research question relating to comparative clauses is to clarify which word(s) the answerers mark intuitively as the head of the word(s) following the comparative conjunction.

2.2 NPs with Participles and Derived Nouns

NPs with sentence-like structures are challenging to parse. Making decisions on how the NP-internal structure should be represented in the dependency grammar framework is a challenging task with no absolute correct solution.

The standard work on Finnish grammar (Hakulinen et al., 2004a) states that if a participle functions as an attribute, it can take an object or an adverbial as a premodifier. The internal structure of an NP with a verb-derived noun as the head of the phrase resembles that of a participial NP. The semantics of the arguments of the head nouns in the following sentences are thus alike.

(5) päivittäin vihanneksia syövä
daily vegetables eating-PR-PRT-ACT
eating vegetables daily

(6) päivittäinen vihannesten syönti
daily vegetables eating-DER
eating vegetables daily

In both examples (5) and (6) the head *syövä/syönti (eating)* takes a direct object: *vihanneksia/vihannesten (vegetables)*. In the participial construction, example (5), the premodifier *päivittäin (daily)* is an adverb directly dependent on the participial head, *syövä (eating)*. In NP (6) the premodifier *päivittäinen (daily)* is an attribute directly dependent on the head noun *syönti (eating)*.

We want to examine whether *vihannesten/vihanneksia (vegetables)* is interpreted as the object in both cases (5) and (6). Traditionally the object has only been seen as the complement of a verb, not of a noun (Hakulinen et al., 2004b).

With the help of an e-query, in which the answerers assign grammatical functions to the premodifiers, we want to examine whether the two constructions, the participial construction, example (5), and the NP with a verb-derived noun as its head, example (6), get analyzed similarly. In addition, we anticipate new insight on the distinction between an adverb and attribute defining a participle or a verb-derived noun.

We extend the research question to cover subjects as well. If a derived noun can take an object as a premodifier, it seems natural that it would analogously be able to take a subject. Consider the following NP:

(7) murhaajan ensimmäinen tappo
murderer's first killingDER
the murderer's first killing

In example (7) the verb-derived noun *tappo (killing)* has a premodifier, *murhaajan (murderer)*. Since the semantics of the sentence cannot be interpreted as the killer being the object of the killing, we want to investigate whether speakers assign *murhaajan* the grammatical function of a subject.

The test we conducted seeks to give new insight on whether the NP's internal grammatical functions are assigned in a parallel manner in participial NPs and NPs with derived nouns. In section 4 we present the results of the experiment.

3 The Experiment

The test is conducted as an online query. We asked Finnish native speakers to answer multiple-choice questions regarding the dependency relations of elliptical verb phrases and sentences and the grammatical function of a participial NP or an NP with a verb-derived head noun. A similar way of using crowdsourcing for collecting linguistic data is described in e.g. Munro et al. (2010).

We presented the respondents a set of ten sentences and asked them to choose the most intuitive answer to the questions from a list of choices. We did not give the respondents the option of inserting a missing element to the elliptical comparative structures because we want to stick to a surface syntax representation.

The 428 answerers are mainly language students and researchers at the University of Helsinki.

They were encouraged to answer the questions swiftly based on their intuition, not according to their knowledge of Finnish grammar. Since the purpose of the query is to find out the users' opinion on the two structures, it does not matter whether their language competence influences their intuitive answers. Most importantly we want to ensure that the future users of the treebank agree with the annotation scheme and that the scheme does not contradict with their language sense.

In the query we collected information about dependency relations (see example question in figure 1) and grammatical functions (figure 2) separately. (For the word-to-word translations, see Appendix A.) To better conceal the aim of the questionnaire, questions on dependency relations alternated with questions on grammatical functions.

Unicafe tarjoaa parempaa ruokaa kuin ennen.
"Unicafe offers better food than before."

What is the head of the word "ennen", i.e. which word is it closest related to?
a. Unicafe
b. tarjoaa
c. parempaa
d. ruokaa
e. kuin

Figure 1: A sample question regarding dependency relations (Sentence 8 in Appendix A.2)

Ojaan pudonnut auto kaivettiin ylös.
"The car that fell into a ditch was dug out."

What is the grammatical function of "ojaan"?
a. predicate
b. subject
c. object
d. adverbial
e. attribute

Figure 2: A sample question regarding grammatical functions (Sentence 1 in Appendix A.1)

Our aim was to estimate if it is possible to get reliable answers to both kinds of questions. The main reason for asking either about dependencies or functions was to not make the questionnaire too time-consuming. Also, we were particularly interested in how the answerers perceive dependency relations in comparative structures on the one hand, and how they assign grammatical functions to complex NPs on the other.

The respondents filled in the questionnaire independently without supervision so we did not monitor the average time taken for answering. We also do not precisely know the background of the answerers, only that most of them are either language students or researchers who heard about the query via mailing lists. The phrasing of the questions did not point the answerers towards dependency grammar but asked the answerers to base their answers purely on intuition.

In order to get a better understanding on the competence of the respondents, the first question in the questionnaire was a control question without elliptical structures or complex NPs. We simply asked the answerers to specify a dependency relation in the following sentence:

Tuuli käy päivisin koulua, ja Vesa työskentelee kotona.
"During the day Tuuli goes to school and Vesa studies at home."

What is the head of the word "kotona", i.e. which word is it closest related to?
a. Tuuli
b. käy
c. päivisin
d. koulua
e. ja
f. Vesa
g. työskentelee

Figure 3: The control question (Sentence 6 in Appendix A.2)

The dependencies in the control question presented in figure 3 are unambiguous so that giving an illogical answer to the question reveals us either that the answerer is not familiar with the notion "head word" or that the answer was marked by accident. The responses to the control question are encouraging: 71% marked *työskentelee (works)* as the head of *kotona (at home)*, and 22% *Vesa*. This leaves us with only 7% illogical answers. Notwithstanding, we regard the results of the questionnaire merely indicative of the answerers intuitive language modeling.

Even though a part of the answers to the control question are not predictable, see example sentence 6 in Appendix A.2, we take all answers into account and do not consider any answers counterintuitive. Still, further research might benefit from narrowing down the results based on the control question.

The experiment presented here is a case study with only 10 questions including one control question. If the experiment would be repeated to cover more phenomena, there should be more questions and different types of control questions. E.g. the elliptical sentences should have a non-elliptical equivalent as a control question to test whether the dependencies are interpreted identically.

4 Results: Modeling the Phenomena

Before determining the syntactic specification for the phenomena, we explore the different ways of modeling them. At this point of the Finnish treebank project, the main goal is not to follow any kind of formalism but to investigate the most natural and semantically informative representation of syntax. Dependency grammar allows for a natural representation of e.g. long-distance relationships because of the non-hierarchical nature of dependency relations (Kübler et al., 2006). At this point we do not try to avoid crossing branches in the dependency trees, since we allow e.g. linking the words of the elliptical comparative sentences to their semantic equivalents in the main clause.

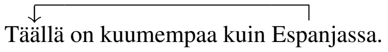
4.1 Elliptical Comparative Structure

The main clause of the comparative clause does not necessarily contain any semantically equivalent word with the word after the subordinating conjunction (see sentence 8 in Appendix A.2). In such a case the most used solution by the answerers is to link the word to the conjunction (55%). The second popular solution is to mark the adjective as the head (20%) and the third popular option for the head is the verb of the main clause (14%).

If the final annotation scheme prefers marking content words as heads, it is worth noticing that 20% of the answerers mark the adjective as the head in a typical elliptical comparative clause with only one word after the conjunction. Also, the conjunction is the most popular choice for the head only when there are no clear semantic or grammatical equivalents in the main clause and no other words in the elliptical clause.

Based on the test, it is intuitively most popular (24%) to link two equivalent words with each other, when the verb of the main clause is *olla (be)*. Example (8) illustrates[4] this solution where the equivalent words, expressions of location, are linked with each other. This tendency to link two compared items to each other supports selecting a representation in which crossing branches are possible.

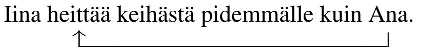
(8)



Täällä on kuumempaa kuin Espanjassa.
Here is warmer than Spain-(ine)
It is warmer here than in Spain.

According to our working hypothesis, the results suggest that when the verb of the main clause is "semantically heavier", the verb is seen as the head more often (33%). This solution is shown in the example (9) where the answerers marked the verb as the head of the elliptical clause even when there is an equivalent in the subject position in the main clause.

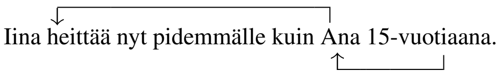
(9)



Iina heittää keihästä pidemmälle kuin Ana.
Iina throws javelin further than Ana
Iina throws the javelin further than Ana.

In the examples above, there is only one word in the comparative clause. When the comparative clause contains an elliptical clause with two or more words, the solutions depend on the interpretation. When there is a primary object of comparison in the comparative clause and the other words are semantically clearly connected to this primary word, it is clearly seen as a head (79%), even if there are equivalent words in the main clause. For example:

(10)



Iina heittää nyt pidemmälle kuin Ana
Iina throws now further than Ana
15-vuotiaana.
15 years old
Iina throws the javelin further now than Ana when she was 15 years old.

When the semantic link between the words of an elliptical comparative clause is not so clear as in example (10), the solutions are so variable that there is no clear conclusion we can draw. Still, based on the answers it is clear that this phenomenon, an elliptical comparative clause, is a real challenge for parsing.

Above we have shown how to treat comparative structures which include elliptical clauses. The comparative sentence can also consist of elliptical phrases, like in the following example[5]:

(11) Matka Tukholmasta Tallinnaan on pidempi
Distance Stockholm-ELA Tallinn-ILL is longer
kuin Riiasta Wieniin.
than Riga-ELA Vienna-ILL

[4] See sentence 7 in Appendix A.2 for the complete answers and the full sentence.

[5] ELA=elative, ILL=illative

The distance from Stockholm to Tallinn is longer than from Riga to Vienna.

The most popular solution (52%) is to connect first part of the elliptical phrase (*Riiasta/from Riga*) to the head of the phrase (*matka/distance*). The latter part of the elliptical phrase (*Wieniin/to Vienna*) was mostly (41%) seen as a dependent of the word *(Riiasta/from Riga)*.

Even though in many cases a semantically heavy word is seen as the head of a comparative clause, throughout the test and in all different kinds of elliptical comparative clauses, the conjunction has always clear support. In all cases, *kuin* is marked as the head of the whole comparative clause by at least 15% of the answerers.

Based on this experiment, we can now roughly sketch the main principles of representing comparative structures intuitively:

- When there is an equivalent sentence element in the main clause, mark it as the head of the dependent in the comparative clause. Link the other parts of the elliptical sentence to this word.

- When there is no equivalent sentence element in the main clause, mark the conjunction as the head of the elliptical comparative clause. When favoring semantically heavier words, mark the adjective as the head as 20% of the answerers do in question 8. (Appendix A.2.).

4.2 Participles and Derived Nouns

The participial NP constructions we wanted the respondents to assign grammatical functions to are the following:

(12) Ojaan pudonnut auto kaivettiin ylös.
Ditch fallenPAST-PRT-ACT car dug up
The car that fell into a ditch was dug out.

(13) Kirkon penkillä itki tekojaan syvästi
Church bench cry deeds deeply
katuva mies.
regrettingPRES-PRT-ACT man
A/the man who deeply regretted his deeds was crying on the church bench.

The primary results of the e-query are assembled in table 1. For conciseness' sake only the three most popular answers are displayed in the table. For the complete results, see Appendix A.1.

The past participles indicate a completed action and have corresponding pluperfect forms. The past participle active form *pudonnut (fallen)* corresponds to a relative clause:

<table>
<tr><td colspan="4">(12) <b>OJAAN PUDONNUT AUTO</b> KAIVETTIIN YLÖS.</td></tr>
<tr><td colspan="4">(13) KIRKON PENKILLÄ ITKI <b>TEKOJAAN SYVÄSTI KATUVA MIES</b>.</td></tr>
<tr><th>Word</th><th>Obj</th><th>Adv</th><th>Attr</th></tr>
<tr><td>ojaan</td><td>47 (11%)</td><td>246 (57%)</td><td>120 (28%)</td></tr>
<tr><td>tekojaan</td><td>250 (58%)</td><td>51 (12%)</td><td>96 (22%)</td></tr>
<tr><td>syvästi</td><td>27 (6%)</td><td>236 (55%)</td><td>158 (37%)</td></tr>
<tr><td colspan="4" align="right">N=428</td></tr>
</table>

Table 1: Grammatical functions of participial NPs

(14) auto, joka oli pudonnut ojaan
car which had fallen into ditch
a/the car which had fallen into a ditch

A participle can get an adverbial modifier (Hakulinen et al., 2004a). In the corresponding relative clause (14) the grammatical function of the premodifier *ojaan (into a ditch)* is adverb. Based on the answers of the e-query, the distinction is not clear in the participial construction. As can be seen from table 1, in fact 57% of the answerers regard *ojaan* an adverb, but as many as 28% consider it an attribute. This might be explained by participles possessing traits of both verbs and adjectives, and the typical modifier of an adjective would be an attribute. Some, 11%, see *ojaan* as an object. This can possibly be explained by the whole NP being the object of the sentence and with semantics: *ojaan* is the target of falling.

In the second participial construction, example (13), we asked the answerers to assign a grammatical function to both of the premodifiers of the participle: *tekojaan (deeds)* and *syvästi (deeply)*. Analogously to the past participle, the present participle *katuva (regretting)* corresponds to a relative clause with a present tense verb.

(15) mies, joka katuu tekojaan syvästi
man who regrets deeds deeply
a/the man who regrets his deeds deeply

Again, the relative clause (15) has clearly distinguishable grammatical functions: *tekojaan* is the direct object of the head verb *katuu*, and *syvästi* is an adverb postmodifying the head.

Analogously, in the participial construction corresponding to the relative clause, 58% of the answerers see *tekojaan* as the object of the sentence. 22% give it the attribute-label, and 12% name it an adverb (see table 1). This indicates that the object premodifier of a participle is a rather straightforward case: a vast majority of the answerers see it as an object.

NPs with a derived noun as their head constitute a similar problem with assigning phrase-internal grammatical functions. Take, for example, the following three sentences from the e-query. We present the most frequent answers in table 2.

(16) Puolet rehtorin ajasta meni oppilaiden
Half principal time went student
ohjaukseen.
guidance

Half of the principal's time was spent on guiding the students.

(17) Päivittäinen vihannesten syönti pitää sinut
Daily vegetables eating keeps you
terveenä.
healthy

Eating vegetables daily keeps you healthy.

(18) Murhaajan ensimmäinen tappo sai paljon
Murderer first kill receive a lot
julkisuutta.
publicity

The murderer's first killing received a lot of publicity.

<table>
<tr><td colspan="5">(16) Puolet rehtorin ajasta meni <b>oppilaiden ohjaukseen</b>.</td></tr>
<tr><td colspan="5">(17) <b>Päivittäinen vihannesten syönti</b> pitää sinut terveenä.</td></tr>
<tr><td colspan="5">(18) <b>Murhaajan ensimmäinen tappo</b> sai paljon julkisuutta.</td></tr>
<tr><th>Word</th><th>Subj</th><th>Obj</th><th>Adv</th><th>Attr</th></tr>
<tr><td>oppilaiden</td><td></td><td>127
(30%)</td><td>43
(10%)</td><td>243
(57%)</td></tr>
<tr><td>vihannesten</td><td>45
(11%)</td><td>130
(30%)</td><td></td><td>218
(51%)</td></tr>
<tr><td>murhaajan</td><td>73
(17%)</td><td></td><td>38
(9%)</td><td>280
(65%)</td></tr>
<tr><td colspan="5">N=428</td></tr>
</table>

Table 2: Grammatical functions of derived NPs

In examples (16) and (17) the NP investigated is in the object position. Both cases reflect a very similar way of intuitive modeling among the respondents: *oppilaiden* and *vihannesten* are given the function of an attribute, 57% and 51%, respectively.

We will now proceed to examine whether a noun can receive an object based on the answerers' intuition. Traditionally only verbs get an object (Hakulinen et al., 2004b), but we want to see if a noun derived from a verb retains this feature of a verb.

The difference between the intuitive response and the object-attribute distinction is clear when comparing the results of the participial NP of sentence (13) and the NPs with a verb-to-noun derivation as the head in sentences (16) and (17). The vast majority (58%) of the respondents label *tekojaan* as an object in (13), whereas only 30% see *oppilaiden* and *vihannesten* in sentences (16) and (17) as the object. This suggests that the verb-to-noun derivations do not possess the traits of a verb, and the traditional definition of the object prevails.

The object-attribute distinction can also be seen from another point of view. As many as 30% of the respondents do in fact think that a noun can receive an object despite the option being excluded by traditional grammars. This suggests that the answerers have a strong semantic way of modeling the phrase alongside with the morphological view.

In sum, intuitive modeling of participial NPs or NPs with a verb-derived head should follow these principles:

- The premodifier of a verb-to-noun derivation is interpreted as an attribute.
- The premodifier of a participial is treated analogously to premodifiers of verbs. It is seen as an object when the verb would take an object, and an adverbial when the verb would have one too.

5 Conclusion

In this paper we have shown that an e-query is a useful tool for collecting information about a treebank's user groups' intuitive interpretations of specific syntactic phenomena. This information is needed to ensure that the syntactic representation used in the treebank does not deviate from its user's language intuition.

Using an e-query for probing for the respondents' intuitive way of modeling syntactic phenomena moves from separate cases to general modeling: A respondent does not need to be consistent with her answers and have one specific answering policy throughout the e-form. Our aim is to collect information about modeling the whole phenomena coherently so these collected opinions are not seen as an unquestionable base for the syntactic model.

Based on this experiment we can also conclude that the variation between the answers results from the fact that these phenomena – the structure of the verb-based NP and the elliptical comparative clause – are semantically ambiguous, and representing them in the dependency grammar framework is not a univocal task. To exclude the possibility of having the same kind of variation in the answers also between other phenomena, we had a control question in the test. The majority of the answers to this question are homogeneous (71%), and the second popular answer (22%) is also semantically valid. This means that 7% of the answers were illogical in a clear-cut case, so at least 7% of the answers should be considered ill-advised. Thus, again we consider the results only as advisory.

Even though the answers to the e-query are varied, some general principles can be made based on our experiment. Interestingly, contradicting the tradition of dependency grammar, where the verb of the main clause is seen as the core of the sentence to which other clauses are related, in some comparative structures the answerers consider e.g. the adjective as the head of the whole comparative clause. This questions the traditional verb-centric modeling of the comparative clauses and suggests perhaps a more informative representation, where the objects of the comparison are more clearly visible.

Based on the number and quality of the answers, an e-query seems to be suitable a suitable method for getting a general view of the users' intuitive way of modeling syntactic phenomena. The large number of the answers also allows for the possibility to eliminate a part of the answers on the grounds of the control question. Before finalizing the syntactic representation of the treebank, we will scrutinize the answers in a more thorough way to receive a more accurate and valid model where the nonsensical answers do not skew the results.

Our experiment shows that the method employed provides new information on how to define the phenomena in the dependency syntactic framework. This information can be used when determining the syntactic specification. The results point towards a way of modeling the syntactic phenomena so that the final syntactic representation used in the treebank does not argue against the view of its users.

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.

References

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.

Sandra Kübler, Jelena Prokić, and Rijksuniversiteit Groningen. 2006. Why is german dependency parsing more reliable than constituent parsing. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT)*, pages 7–18.

Robert Munro, Steven Bethard, Steven Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics.

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

Atro Voutilainen and Krister Lindén. 2011. Designing a dependency representation and grammar definition corpus for finnish. In *Proceedings of III Congreso Internacional de Lingüística de Corpus (CILC 2011) (upcoming)*.

A Complete Results

The total number of answers is 428.
Percentages are rounded to the nearest whole number.

A.1 NP Constructions

Word	Predicate	Subject	Object	Adverb	Attribute	NA
1. **OJAAN PUDONNUT AUTO** KAIVETTIIN YLÖS. "The car that fell into a ditch was dug out."						
ojaan into a ditch	5 (2%)	5 (2%)	47 (11%)	246 (57%)	120 (28%)	5 (2%)
2. PUOLET REHTORIN AJASTA MENI **OPPILAIDEN OHJAUKSEEN.** "Half of the principal's time was spent on guiding the students."						
oppilaiden students'	3 (1%)	6 (1%)	127 (30%)	43 (10%)	243 (57%)	6 (1%)
3. **PÄIVITTÄINEN VIHANNESTEN SYÖNTI** PITÄÄ SINUT TERVEENÄ. "Eating vegetables daily keeps you healthy."						
vihannesten vegetables-GEN	3 (1%)	45 (11%)	130 (30%)	22 (5%)	218 (51%)	3 (2%)
4. **MURHAAJAN ENSIMMÄINEN TAPPO** SAI PALJON JULKISUUTTA. "The murderer's first killing received a lot of publicity."						
murhaajan murderer's	2 (0%)	73 (17%)	14 (3%)	38 (9%)	280 (65%)	21 (5%)
5. KIRKON PENKILLÄ ITKI **TEKOJAAN SYVÄSTI KATUVA MIES.** "The man who deeply regretted his deeds was crying on the church bench."						
tekojaan deeds-PAR	1 (0%)	7 (2%)	250 (58%)	51 (12%)	96 (22%)	23 (5%)
PAR=PARTITIVE, GEN=GENITIVE						

A.2 Comparative Constructions

The following tables show what is seen as the head of the word in italics:

6. TUULI KÄY PÄIVISIN KOULUA, JA VESA OPISKELEE KOTONA. "During the day Tuuli goes to school and Vesa studies at home."						
Word	**Tuuli** Tuuli	**käy** goes	**päivisin** daily	**koulua** to school	**Vesa** Vesa	**opiskelee** studies
kotona at home	2 (0%)	14 (3%)	6 (1%)	6 (1%)	96 (22%)	304 (71%)

7. TÄÄLLÄ ON KUUMEMPAA KUIN TURISTEILLA KESÄLLÄ ESPANJASSA. "It is hotter here than what tourists experience in Spain during the summer."								
Word	**Täällä** Here	**on** is	**kuumempaa** hotter	**kuin** than	**turisteilla** tourists-ADE	**kesällä** in the summer	**Espanjassa** in Spain	**NA**
turisteilla tourists-ADE	25 (6%)	46 (11%)	59 (14%)	105 (25%)	- -	36 (8%)	126 (29%)	31 (7%)
kesällä in the summer	26 (6%)	30 (7%)	50 (12%)	32 (7%)	83 (19%)	- -	175 (41%)	32 (7%)
Espanjassa in Spain	103 (24%)	29 (7%)	52 (12%)	64 (15%)	84 (20%)	63 (15%)	- -	33 (8%)
ADE=ADESSIVE								

8. Unicafe tarjoaa parempaa ruokaa kuin ennen.
“Unicafe offers better food than before.”

Word	**unicafe**	**tarjoaa**	**parempaa**	**ruokaa**	**kuin**	**NA**
	Unicafe	offers	better	food	than	
ennen	10	59	87	17	234	21
before	(2%)	(14%)	(20%)	(4%)	(55%)	(5%)

9. Iina heittää keihästä jo nyt pidemmälle kuin Ana 15-vuotiaana.
“Iina throws the javelin further already now than Ana when she was 15 years old.”

Word	**Iina**	**heittää**	**keihästä**	**jo**	**nyt**	**pidemmälle**	**kuin**	**Ana**	**15-vuotiaana**	**NA**
	Iina	throws	javelin	already	now	further	than	Ana	15 years-ESS	
Ana	59	142	16	0	1	38	129	-	31	12
	(14%)	(33%)	(4%)	(0%)	(0%)	(9%)	(30%)	-	(7%)	(3%)
15-vuotiaana	7	21	5	5	21	6	15	338	-	10
15 years-ESS	(2%)	(5%)	(1%)	(1%)	(5%)	(1%)	(4%)	(79%)	-	(2%)

ESS=ESSIVE

10. Matka Tukholmasta Tallinnaan on pidempi kuin Riiasta Wieniin.
The distance from Stockholm to Tallinn is longer than from Riga to Vienna.

Word	**Matka**	**Tukholmasta**	**Tallinnaan**	**on**	**pidempi**	**kuin**	**Riiasta**	**Wieniin**	**NA**
	Distance	Stockholm-ELA	Tallinn-ILL	is	longer	than	Riga-ELA	Vienna-ILL	
Riiasta	222	41	1	5	27	67	-	11	17
Riga-ELA	(52%)	(10%)	(0%)	(1%)	(6%)	(16%)	-	(48%)	(4%)
Wieniin	138	3	40	2	26	22	176	-	21
Vienna-ILL	(32%)	(1%)	(9%)	(0%)	(6%)	(5%)	(41%)	-	(5%)

ELA=ELATIVE, ILL=ILLATIVE