A Double-Blind Experiment On Interannotator Agreement: The Case Of Dependency Syntax And Finnish

Atro Voutilainen and Tanja Purtonen
Department of Modern Languages
University of Helsinki
atro.voutilainen@helsinki.fi tanja.purtonen@helsinki.fi

Abstract

Manually performed treebanking is an expensive effort compared with automatic annotation. In return, manual treebanking is generally believed to provide higher-quality/value syntactic annotation than automatic methods. Unfortunately, there is little or no empirical evidence for or against this belief, though arguments have been voiced for the high degree of subjectivity in other levels of linguistic analysis (e.g. morphological annotation). We report a double-blind annotation experiment at the level of dependency syntax, using a small Finnish corpus as the analysis data. The results suggest that an interannotator agreement can be reached as a result of reviews and negotiations that is much higher than the corresponding labelled attachment scores (LAS) reported for state-of-the-art dependency parsers.

1 Introduction

There is ongoing effort in many countries on treebank annotation to support linguistic research, statistical language modelling and other tasks (Haverinen et al., 2009; Kromann, 2003; Marcus et al., 1993; Mikulova et al., 2006; Nivre et al., 2006). Treebanks are usually text collections with (tens of) thousands of sentences annotated according to a dependency syntactic or phrase structure representation documented as an annotator's manual.

Annotation can be made automatically or manually. Treebanks created with a parser can be very large, because automatic parsing is a fast and inexpensive operation. Manual annotation is slower: the creation of manually annotated treebanks tends to take many years, as reported by several presenters at a recent CLARA Treecourse in Prague (Dec. 2010). Still, treebanks annotated by hand are considered more valuable, because manual annotation is believed to result in higher accuracy.

Unfortunately, it is difficult to find empirical evidence to support or question this belief in annotation accuracy benefits. At other levels of linguistic analysis, the so-called "double blind experiment" has been used for measuring interannotator agreement (Kilgarriff, 1999; Voutilainen, 1999). At the syntactic level, we are not aware of such experiments.

Without relevant empirical data, one can question the investment needed for manually annotating a treebank, e.g. by using the following conjecture: if human annotators can after negotiations disagree about the correct analysis even in 5% of words at the POS level (Church, 1992), annotator disagreement in the (assumedly) more complex task of syntactic annotation is likely to be so much higher, that there might be no actual advantage in annotation quality, when comparing a manually annotated treebank with an automatically annotated one.

In this paper, we report a small-scale double-blind experiment on dependency syntactic annotation using Finnish-language text as empirical data. We provide interannotator agreement figures before and after the negotiation phase, as well as more observations on types and apparent reasons for annotation differences.

Our experiment suggests that with a carefully documented linguistic representation, human annotators can agree on a syntactic analysis to a much higher degree (jointly achieving labelled attachment scores of close to 99%) than what even the best syntactic analysers are reported to reach (80-90% LAS scores). – How much of the high agreement rate can be generalised to other dependency syntactic annotation models and practices remains a topic for future research.

Next, we outline the key characteristics of

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 319–322

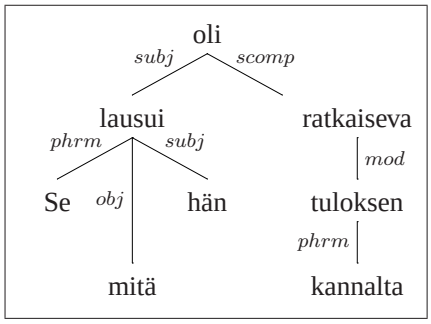
the syntactic representation and its specification. Then we describe the double-blind experiment and empirical data. This is followed by quantitative results and general observations.

2 Syntactic Representation And Its Specification

The syntactic representation used in this experiment can be characterised as follows:

- each word has a unique syntactic head;
- the representation is surface-syntactic (no empty categories postulated);
- dependency structures can be non-projective (Finnish as a free word-order language has unbounded dependencies);
- grammatical markers or attributes (e.g. articles, quantifiers, other modifiers, prepositions, postpositions, conjunctions, auxiliaries) are treated as dependents; semantically "heavy" words are preferred as heads;
- to each dependency relation, a syntactic function is attached;
- the syntactic function palette contains 15 basic functions (e.g. auxiliary, phrasal particle, subject, object, vocative).

Here is a sample syntactic analysis where what is normally called formal subject "se" (english "it") is analysed as a phrase marker for the actual subject clause; note also phrase marker analysis of the postposition "kannalta".



Se [it] mitä [what] hän [s/he] lausui [said] oli [was] tuloksen [result.Gen] kannalta [prom-the-point-of.Postp] ratkaiseva [decisive]. ("What s/he said was decisive for the result")

The syntactic specification is based on an initial draft completed when annotating some 19,000 hand-picked corpus-based sentences used as examples in a descriptive grammar of Finnish (Hakulinen et al., 2004). Specifying the grammatical representation as an annotator's manual was expected to be more successful because the inventory of grammatical constructions is readily available in the form of such a "grammar definition corpus". The manual will be published online at `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank`.

3 Test Arrangements

The double-blind experiment was conducted as follows. Firstly, two trained annotators independently marked the function and the dependency of every word in their own corpus version, presented in spreadsheet form similar to CONLL-X (`http://nextens.uvt.nl/~conll/#dataformat`). The text was automatically tokenised and morphologically analysed, and the annotators were aware that there can be errors in morphological analysis (but no corrections to morphology were made). The annotators were encouraged to consult the annotator's manual, the syntactically annotated grammar definition corpus and the descriptive Finnish grammar (`http://kaino.kotus.fi/visk/etusivu.php`) from which the example sentences were extracted.

Secondly, these manually annotated versions of the text were automatically compared with each other. Words with a different analysis were marked with a symbol "LOOK", which was added also to some random words to minimize the risk of only guessing the other annotator's analysis. At this point (round 1), the annotators were not aware of each other's answers, and independently made the corrections to their own corpus versions.

Thirdly, the reanalysed texts were automatically compared with each other, and words with a different analysis were re-marked. At this point (round 2), the annotators saw each other's answers, and they negotiated about the disagreements, and documented their negotiations. On the basis of the negotiations, the differences between analyses appear to result from five main reasons ("D:a–e" in tables 2 and 3):

- (D:a) Lack of attention.

- (D:b) Incomplete specification in the manual. After negotiating, the annotators could find a common solution (to be added to the manual).

- (D:c) Incomplete specification of the manual, but after negotiations the annotators agreed that a separate study is needed to cover the phenomenon. So, at this stage category could not be analysed consensually and unambiguously.

- (D:d) Real ambiguity.

- (D:e) Domino effect.

The routine was successively applied to each text in the test corpus.

The test corpus consisted of three texts from three genres, totalling 2039 words and 176 sentences:

- fiction: 561 words of a novel by Jostein Gardner ("Sophie's world");

- news: 694 words from online editions of "Helsingin Sanomat" and "Tietoviikko" (11.1.2011);

- Wikipedia: 784 words from three Wikipedia articles on geography and history.

4 Results

The results from the double-blind experiment are presented in Tables 1–3.

Corpus and stage	Agreement rate
fiction (1)	89.7% (503/561)
fiction (2)	92.6% (519/561)
fiction (3)	98.6% (553/561)
news (1)	90.8% (630/694)
news (2)	96.3% (668/694)
news (3)	98.7% (685/694)
wikipedia (1)	88.9% (697/784)
wikipedia (2)	94.8% (743/784)
wikipedia (3)	99.2% (778/784)

Table 1: Word-level interannotator agreement rates for dependency relation+function analysis before review (1), before negotiation (2), after negotiation (3).

Data	D:a	D:b	D:c	D:d	D:e	Total
Fiction (2)	8	2	4	4	14	32
News (2)	4	2	6	3	7	22
Wiki (2)	9	9	2	2	16	38
Total (2)	21	13	12	9	37	92

Table 2: Classification of differences in dependency relation analysis.

Data	D:a	D:b	D:c	D:d	D:e	Total
Fiction (2)	6	8	4	3	14	35
News (2)	5	4	6	1	7	23
Wiki (2)	2	10	4	-	16	32
Total (2)	13	22	14	4	37	90

Table 3: Classification of differences in dependency function analysis.

The following two tables show the different analyses classified to the differences in the dependency relation analysis (table 2) and in dependency function analysis (table 3).

The disagreement rate diminished clearly between rounds 1 and 2 and 3. Still, many clerical errors (due to inattention) persisted even at stage 2. Syntactic annotation with a spreadsheet may be more error-prone than with a tree editor.

5 Discussion

Some general points are in order. Firstly, the grammar corpus is created from the example sentences in (Hakulinen et al., 2004). The descriptive grammar appears to focus on traditional (theoretically interesting) types of syntactic phenomena, like common vs special clause types. Much less attention seems to be given e.g. to different types of names and titles and their combinations, to quantitative expressions, and to expressions with numerals or other fixed-form material. In this experiment, the annotators were able to analyse even syntactically complex and long sentences (e.g. many embedded sentences) remarkably consistently, but the annotations repeatedly differed in the case of "local" expressions such as temporal or areal expressions, which were not covered in the annotator's manual.

The annotation differences between genres were remarkable. It may result from the fact that the news articles are mostly written using standard language, but in the fiction text, there are many el-

liptical sentences. A difference between analyses, especially in the elliptical cases, often causes the domino effect, and in the test corpus, 41% of all differences in annotation are caused by the domino effect at word level.

The test corpus consisted of continuous text, but the annotated 19,000-sentence grammar definition corpus contains mostly isolated sentences. To account for elliptical constructions (and other supersentential phenomena), the grammar/manual definition phase should benefit from continuous corpus texts, in addition to systematic grammar corpus sentences, to enable a more informed analysis.

In this experiment, the double-blind-method was used for estimating, to what extent interannotator agreement can be reached; and the aim was not to avoid differences in annotation. Still, many of the (initial) differences in syntactic annotation can probably be avoided by providing also a visual interface to the annotators, who in this experiment worked with tabular spreadsheet format only. Also, the annotator's manual needs a fair supply of annotated example sentences to concretise the more abstract descriptive statements on some particular category.

To conclude: in this paper, we have documented a double-blind experiment on syntactic annotation to provide an initial understanding (based on limited empirical data) on what level of annotation consistency can be reached by human annotators at the level of syntactic analysis. Our experiment shows that a much higher agreement rate (around 99%) on the correct syntactic annotation can be reached than is reported as the corresponding word-level labelled attachment score (LAS) for state-of-the-art dependency parsers (close to 90% LAS for English; 70–80% LAS for other languages with richer morphology and less rigid word order).

Acknowledgments

We wish to thank members of the FIN-CLARIN HFST team for their support of this work, in particular Sam Hardwick for programming support. We also gratefully acknowledge constructive reviews of the three anonymous NODALIDA referees.

References

Kenneth W. Church. 1992. Current Practice in Part of Speech Tagging and Suggestions for the Future. In Simmons (ed.), *Sbornik praci: In Honor of Henry Kucera*. Michigan Slavic Studies. Michigan. 13–48.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen and Irja Alho. 2004. *Iso suomen kielioppi* [Large Finnish Grammar]. Helsinki: Suomalaisen Kirjallisuuden Seura. Online version: http://scripta.kotus.fi/visk URN:ISBN:978-952-5446-35-7.

Katri Haverinen, Filip Ginter, Veronica Laippala, Tapio Viljanen and Tapio Salakoski. 2009. Dependency Annotation of Wikipedia: First Steps towards a Finnish Treebank. *Proceedings of The Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, pp. 95–105.

Adam Kilgarriff. 1999. 95% Replicability for Manual Word Sense Tagging. *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

Matthias Kromann. 2003. The Danish Dependency Treebank and the underlying linguistic theory. *Proc. of the TLT 2003*.

Mitchell Marcus, Beatrice Santorini and Mary Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2).

Marie Mikulova, Alevtina Bemova, Jan Hajic, Eva Hajicova, Jiri Havelka, Veronika Kolarova, Lucie Kucova, Marketa Lopatkova, Petr Pajas, Jarmila Panevova, Magda Razimova, Petr Sgall, Jan Stepanek, Zdenka Uresova, Katerina Vesela, and Zdenek Zabokrtsky. 2006. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report 30, UFAL MFF UK, Prague, Czech Rep.

Joakim Nivre, Jens Nilsson and Johan Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*.

Atro Voutilainen. 1999. An experiment on the upper bound of interjudge agreement: the case of tagging. *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955