Teaching treebanking

Volk, Martin ; Gusafson-Capková, S ; Hagstrand, D ; Uibo, H

Teaching Treebanking

by

Martin Volk, Sofia Gustafson-Capková and David Hagstrand (Stockholm University)
Heli Uibo (Tartu University)

1. Introduction

Treebanks have become valuable resources in natural language processing (NLP) in recent years (Abeillé, 2003). A treebank is a collection of syntactically annotated sentences in which the annotation has been manually checked. The name derives from the fact that syntactic descriptions of sentences often come in the form of tree structures, in particular constituent trees. But treebank annotation has also been done in the framework of dependency grammar and recent annotation has also exceeded syntax towards semantic features such as predicate-argument structures or word senses.

A treebank can serve as training corpus for natural language parsers, as repository for linguistic research, or as evaluation corpus for NLP systems. In particular it has been shown (Manning and Schütze, 1999) that natural language parsers which are trained on treebanks (rather than being based on hand-crafted rules and preferences) are

more robust and thus more successful for practical applications. Treebanks have also become a necessary resource for many research activities in NLP.

But while treebanking has proven to be of high importance for NLP, instruction in this field has been largely missing. Students have learned about treebanks within syntax courses or within general courses on natural language processing or on corpus linguistics. But treebanking is a time-consuming and demanding activity and therefore specific training is essential. We therefore organized an intensive course on "Treebanks: Formats, Tools and Usage" for PhD students in Language Technology from the Nordic countries. The course took place at Stockholm University in March 2004 and was attended by more than 20 participants from Denmark, Estonia, Finland, Iceland, Norway, and Sweden. Due to its success it was repeated in a reduced manner for undergraduate students as part of the summer school on "Empirical Methods in NLP" at Tartu University in August 2004. Again the course was attended by more than 20 participants, this time coming from the Baltic countries, Russia and Ukraine.

Participants at the Stockholm University treebanking course

The courses introduced the processes involved in creating and exploiting treebanks. They gave an overview of the annotation formats in different treebanks (e.g. the English Penn Treebank, the Swedish Treebank SynTag, the German TIGER Treebank, the Danish Dependency Treebank etc.). And they demonstrated the most important tools used for the creation of treebanks (tree editors), for consistency checking in treebanks and for treebank searches. The general goal was to offer the students hands-on experience in treebanking so that they will be able to participate in treebank projects for their own languages or even help initiating such projects.

This report documents these two treebank courses. We summarize the topics covered in the lectures and comment on what we find most central. We describe how we set up the practical exercises and give recommendations for future practice sessions. We also share our observations about how to adapt the material from a treebank course on the PhD level to an undergraduate course. We conclude with some remarks on supporting activities that accompanied both courses and an outlook.

2. Course format

The PhD course in Stockholm (2 credit points) consisted of reading seven research papers before the course, 15 hours of lectures and 12 (+ 3) hours of computer labs. Students who were willing to work on an individual research project could obtain 5 credit points. The Treebank course in Tartu (1 credit point) included 8 hours of lectures and 8 hours of hands-on sessions. The core part of the lectures was similar in Stockholm and Tartu. The course in Stockholm additionally contained some guest lectures on different treebank-related topics (training of a parser, spoken language treebanks and the Danish Dependency Treebank), while the course in Tartu provided new research results about parallel treebanks. For the Tartu course, three out of the four practical sessions designed for the Stockholm course were used in a reduced format to address the undergraduate level.

3. The Treebank Lectures

Our ambition with the lectures in the treebank course was to cover the most important aspects of treebanking, from the sampling and collection, over annotation and search into future enrichments of treebanks. In addition to this general overview of treebank construction, we wanted to provide insights into previous and ongoing work with specific treebanks.

Composition of the lecture series

To fulfill the aims defined above, we let one set of lectures cover more general aspects, and to satisfy the requirements of more specific aspects, external speakers with experience from specific projects in the treebanking field were invited. The lectures dealing with general aspects were taught by staff from Stockholm University and covered the following topics:

- Corpus collection and preparation
- Treebank tools and treebank search
- Treebank maintenance
- Treebanks and discourse information
  The invited speakers' talks covered the topics of:
- Training a parser for Swedish on a treebank (Beata Megyesi, KTH Stockholm)
- Spoken language treebanks (Erhard Hinrichs, University of Tübingen).
- The Danish Dependency Treebank (Matthias Trautner Kromann, Copenhagen Business School)

An introduction plus a closing talk with a glance into the future, completed the lecture series. The lectures were scheduled in the mornings while the afternoons were reserved for practical exercises (see section 3, Practical Treebank Exercises).

The lectures covering general aspects of treebanking were distributed over four topics and are briefly described below.

1. Corpus collection and preparation

In this lecture we covered questions concerning the sampling and balancing of a corpus as well as how requirements on the sampling can differ with regard to the intended purpose. Examples of the requirements of the SUC corpus (Ejerhed et al.) and its sampling dimensions were shown and discussed. In addition the choice of format for a corpus was touched upon.

2. Treebank tools and searching

Two lectures were addressing this topic, covering different categories of tools useful in the construction (graph editors and disambiguators, such as e.g. Annotate [Anno])

as well as search tools, such as e.g. TIGERSearch [Tiger], developed for the German TIGER Treebank and TGrep2 developed for search in the Penn Treebank (PTB) [Penn].

3. Treebank maintenance

This lecture addressed the topic of how to maintain the corpus after finishing a first version. It was discussed how to handle bug reports, how to distribute, update and enrich a treebank, and how to keep the consistency of a treebank. The legal issues were also brought up. The close connection between the construction (including the selection of format) and future addition of new information was stressed.

4. Treebanks and discourse information

In this lecture we showed how existing treebanks have been enriched with discourse information. Examples were drawn from the RST Discourse Treebank (Marcu et al., 1999), which is annotated with rhetorical relations and the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004) annotated with coherence relations from a discourse parser based on DL-TAG. Examples of tools for discourse annotation, such as RST-tool and Wordfreak were given, and the lecture closed with a discussion of different approaches to discourse structure.

In addition to the lectures covering treebanking topics in general, the three invited speakers each gave one talk in the area of their own specific experiences with treebanking.

Beáta Megyesi from KTH, Stockholm, gave a talk on her experiments on training a shallow parser for Swedish (Megyesi, 2002), focusing on parsing methods and problems related to parsing. Erhard Hinrichs from University of Tübingen presented experiences from his work with the development of the Tübingen Treebank of German Spontaneous Speech (TüBa-D/S) (Stegmann, 2000), and discussed problems specific for spoken language and dialogue. And Matthias T. Kromann, CBS Copenhagen, gave a talk on his work with the Danish Dependency Treebank (Kromann, 2003), which consists of a part of the Danish PAROLE corpus analyzed with the dependency-based formalism Discontinuous Grammar. Most weight was put on the discussion of the analysis and formalism.

The invited speakers, showing practical examples from their own work with treebank development constituted a natural link between the more theoretical underpinnings in the rest of the lectures and the students' hands-on work during the practical sessions. The reactions of the course participants indicated that the selection of lecture topics for the course was rather well picked. Of course there are topics that could be considered as more "core treebanking", such as e.g. the actual annotation of trees including the grammatical formalism and the parsing and annotation tools. However, in constructing a treebank, the quality and usability is dependent also on insights gained in the field of corpus construction in general. That is why the topics, such as e.g. sampling, formats and legal issues, have their given place also in a course directed specifically towards treebanking.

4. Practical Treebank Exercises

This section describes the different exercises designed for the course, along with descriptions of what was required to set them up in terms of software, data resources and other preparations. In all, we constructed three exercises covering different aspects of hands-on treebank work.

General setup

The majority of work for the preparation of the hands-on sessions was done at Stockholm University. In Tartu most of the resources could be re-used.

In Stockholm all software needed for the exercises was installed on a Linux server, and each student was given an account on the server.

In Tartu some of the programs and data were installed on a remote server (treebank annotation tool Annotate and the treebanks imported into it as MySQL databases), others locally on the Linux computers (the searchable treebank, TGrep2, TIGER-Search, Perl scripts). The students of Tartu University used their existing accounts on the university server. For other students guest accounts were set up, such that they could log in both to Linux computers and to the remote server. The number of computers in the computer labs in both Stockholm and Tartu was less than number of participants, which we actually saw as an advantage since it made the students work in teams.

Exercise 1 – Annotation

The goal of the first exercise was to illustrate the difficulties of making consistent decisions in the process of annotating a treebank, by letting the students annotate 50 English sentences according to the Penn Treebank Guidelines [Penn-guide].

For this exercise the tree editor Annotate [Anno] was used, and we trained the built-in PoS-tagger and chunker on 90% of the Wall Street Journal (WSJ) part of the Penn Treebank. Each student was assigned a blank corpus (i.e. without annotation) consisting of 15 sentences taken from the remaining 10% of WSJ, plus 35 raw sentences collected from the online edition of Atlanta Journal-Constitution [AJC]. The first 15 sentences, serving as practicing material, were identical across all groups and were also available in a separate read-only Annotate corpus with all annotation kept intact, for use as reference. As soon as the students felt comfortable with the annotation process, they could move on to the remaining 35 sentences. To complete the task, the students were supposed to hand in 15 fully annotated sentences from the AJC set, along with a description of their annotation decisions. At the Stockholm course some students worked on treebanks of their own language (Icelandic, Estonian, Bulgarian, Amharic). It made the task more rewarding for the students as it had practical value for the Sofie Parallel Treebank which is being developed in the context of the Nordic Treebank Network [NTN]. And it is also easier to analyze sentences in one's mother tongue. But this approach also had a disadvantage, as it was not possible to use a tagger and chunker for semi-automatic annotation, as could be done for English.

Comments on exercise 1

Annotate is a good tool for annotating treebanks, however it has its drawbacks. Firstly, the administration of corpora and users allowed to edit or view these are not straightforward. All Annotate's data is kept in MySQL databases, and much of the administrative work has to be done outside of Annotate, dealing with the databases themselves. Common tasks such as adding a new corpus or a new user turns out to be quite complex and time consuming. Secondly, the user interface of Annotate is sometimes confusing, and it takes some time of practice until one can use it effortlessly. Despite these drawbacks, our impression is that Annotate's strengths in the end wins over its weaknesses, making it recommendable for others to use both in training and actual treebanking work.

Exercise 2 – Extracting information from a Treebank

Through this exercise we wanted to let the students experience what it takes to extract information from a treebank. We decided to use TGrep2 [Tgrep] as the extraction tool, and parts of a large treebank as the data source. The students were given a brief introduction to the TGrep2 query language, and were then asked to perform a series of tasks of varying complexity, such as finding instances of particular structures in the treebank. To complete the tasks, they were supposed to deliver their extracted instances from the treebank along with the TGrep2 commands that extracted them.

Comments on exercise 2

As expected, the students with experience from tools like egrep and similar had a steeper learning curve for the TGrep2 query language than the students lacking these skills. However, even the groups with the least computing experience managed to complete the tasks within the timeframe.

No doubt, TGrep2 is a powerful search tool. However, its main disadvantage is that it is a command line tool. All output is text based, making the results quite cumbersome to interpret. Here TIGERSearch offers a good alternative with the capability of illustrating query results as actual trees.

Exercise 3 – Evaluating a Chunker against a Treebank

The goal of this exercise was to let the students experience what it takes to write chunker rules and to evaluate the resulting chunker against a treebank.

The students were provided with a rule-based chunker written in Perl (Volk, 2001). The task was to write rules for the chunker in order to identify as many phrases as possible in a corpus consisting of 500 sentences from a treebank, stripped of all annotation except for POS-tags. After applying the rules on the corpus, they should evaluate the results against a gold standard – the same 500 sentences with all annotation intact. The evaluation was done automatically by a Perl program which reported recall and precision figures, both overall and per phrase.

Comments on exercise 3

The exercise turned out well, however some students found it quite hard to keep control over the rules as their number grew. This was much due to the fact that the chunker program in its present version gives little feedback, which makes identifying e.g. overgenerating rules difficult. For future use of this exercise we will improve the feedback from both the chunker as well as the evaluation programs.

This exercise showed the biggest discrepancy between groups with respect to their experience of working with Linux. Some groups struggled with the basic steps of handling the different Perl programs, whereas others made use of their wider knowledge by e.g. autogenerating chunker rules directly from the treebank, using tools such as TGrep2, egrep and sed.

Summary on the exercises

In general, all exercises were successful, both from a technical and didactic view. As previously mentioned, the students greatly appreciated the mix of theory and practical work during the course. We were also content to see that the students lacking previous programming skills managed to complete all tasks without major problems, which we had our concerns about when designing the exercises.

Another useful exercise would be to use a treebank for parser training and evaluation, letting the students use increasingly sized parts from the treebank as training data and then to observe the effects on the parsing quality.

As a concluding remark we would like to stress the importance of preparing exercises like this in good time before the course. The ideal situation would be to have a test group trying out all exercises in advance, in order to find the pitfalls. These could be everything from failing software to badly written instructions.

5. Teaching treebanking to graduate and undergraduate students

Students' backgrounds

The treebank course at Stockholm University was announced as a PhD course and all the participants were graduate students from the Nordic countries. But the students' background and aims were still quite different, as the research in corpora, syntactic

parsing and data-driven methods have longer traditions in some of the universities, but for students from some other sites the concept of treebank was quite new.

The intended audience of the summer school „Empirical methods in NLP“ in Tartu were advanced undergraduate students specializing in computer science, linguistics, or computational linguistics and no strict prerequisite was applied. The diversity of students' backgrounds was quite high, including both undergraduates and graduates with stronger background in either linguistics or computer science, some young people working at software companies and some university teachers of linguistics.

CL curricula in different universities over the world can be divided in two – ones where students start with mathematics and computer science and the others where they start from linguistics. And there are a lot of universities where no such speciality as computational linguistics exist. There were representatives from all three kinds of universities among the participants of the Tartu summer school.

Treebanking with undergraduates

Taking into account the expected difference in computer skills of students at the Tartu summer school, we have composed more detailed technical instructions for this course (step-by-step instructions for command-line commands, quick links to the most essential programs etc). We also prepared the working environment as much as possible (wrote shell scripts which set environment variables when logging in) letting the students concentrate on the contents of the tasks.

The first exercise (treebank annotation with Annotate) was purely linguistic. It assumed a good knowledge of English syntax and phrase structure grammar. However, it was the most difficult task for most of the students at summer school. It was not surprising that students who had a computer science background had not much knowledge about syntactic theories. But it also turned out that many of the linguistics students from the previous Soviet countries had not been taught a classical syntax course, which includes the principles of constituency grammar as it is taught in Western universities. Rather they had some knowledge of dependency grammar. The students whose main subject was the Estonian language had studied Estonian syntax but had not studied English syntax systematically enough. During this exercise the summer school students often needed technical assistance, as the user interface of the program Annotate is quite tricky. It takes some time to get used to the functions of left and right mouse button clicks for node grouping and ungrouping.

Although the course in Tartu did not require computer programming, writing complex tgrep queries which combine different operations and use parentheses is quite similar to writing complex conditional statements in programming. To write appropriate queries, the student should have at least some mathematical literacy, knowledge of set theory and mathematical logics. Therefore the treebank searching task was quite difficult for undergraduate students with a background in linguistics. Our teaching experience shows that it is very important to study mathematical subjects constantly, to keep the mind "trained" for mathematical thinking.

For the chunker training students had to create rules which could be turned into the precision and recall values by sequential application of three Perl programs. Most students agreed it was an exciting task. To encourage the students even more, a competition was announced: who will get the best values of precision and recall? In Stockholm the team of students from the University of Oslo won the game. In Tartu the Latvian students were most successful. Both graduates and undergraduates mostly understood the syntax of the rules, but for many students the creation of an ordered rule set was a difficult task. To some undergraduates the notions of precision and recall were not familiar.

The PhD course on treebanks in Stockholm was well tailored for the PhD students whose research topics were related to syntax, syntactically annotated corpora or data-driven methods (this was the intended target group). At this course the difference in preparation was well smoothed by working in small groups during the hands-on sessions. In Tartu three practical assignments on four lab sessions enabled advanced students to help the slower students to finish their tasks.

However, there were some difficulties in adapting the PhD course in Stockholm for the undergraduate course in Tartu, as many of the summer school students were lacking a background in syntax. In principle, the course is well built and can be adapted to undergraduates but in that case certain prerequisites should be set – the list of courses passed or equivalent skills and knowledge acquired (e.g. syntactic theories, especially phrase structure grammar, with examples from English grammar; mathematical logics and set theory; computational linguistics; computer handling – work on the Unix/Linux command line).
All the students admitted that the lectures were very interesting, and the hands-on sessions have added double value to the course. A lot of students had the opportunity to work on a real treebank for the first time.

6. Supporting activities

The PhD course was followed by projects giving students the opportunity to earn additional credit points. The following projects were undertaken:

Two projects were about parallel treebanks, both on the topic of transferring information from a treebank in one language (e.g. EN) to a parallel language (e.g. Amharic). Atelach Alemu (Stockholm) worked on "Projecting Dependency Parses - English to Amharic". She has parsed English sentences from the novel “Sofie’s World” and wrote a program to transfer the information to Amharic, an Ethiopian language. Her conclusions were rather negative since the two languages are so different. However, Svetoslav Marinov (Skövde) experimented with a similar approach for the "(Semi-)Automatic transfer of syntactic information" from Swedish to Bulgarian. He transferred the dependency information computed for the Swedish version of “Sofie’s World” by the Växjö group to Bulgarian. And his recall and precision values were encouraging.

Johan Hall and Jens Nilsson (Växjö University) worked on "Converting dependency treebanks to MALT-XML" including the conversion of non-projective to projective dependency structures. Kaarel Kaljurand (Tartu University) worked on "Checking treebank consistency" and evaluated his program on parts of the German NEGRA Treebank. Henrik Oxhammar and Hans Hjelm (Stockholm University) worked on "Guidelines for Named Entity Markup in the Treebank Editor ANNOTATE" as a basis for studying the structures of product names.

The Treebank course in Stockholm was accompanied by a panel discussion on the use of linguistic theories in natural language processing. Jussi Karlgren from SICS (Stockholm) argued that linguistic theories often obscure progress in language technology rather than supporting it. Professors Erhard Hinrichs (Computational Linguistics, Tübingen) and Östen Dahl (General Linguistics, Stockholm) added their perspective on the topic. The lively discussion was moderated by Rickard Domeij.

Participants following the panel discussion

Both the PhD course in Stockholm and the Summer School in Tartu were rounded off by social activities that contributed to the pleasant atmosphere and the friendly cooperation between all the participants (reception and dinner in Stockholm, excursion to Southern Estonia and farewell dinner in Tartu).

Excursion of the summer school participants to Southern Estonia

7. Conclusions

The treebank courses in Stockholm and Tartu received excellent grades from the participating students for their balance between teaching and practice sessions. The enthusiastic students made them a pleasure ride also for the teaching staff.

Future courses might profit from more focused reading material. A textbook on treebanks is still missing. But upcoming courses will also have to incorporate new developments in the field as for example tree fragmentation as a basis for data-oriented parsing, the issue of parallel treebanks including special tools for sub-sentential alignment, or more innovative annotation in semantics and discourse.

8. References

Anne Abeillé (ed.): *Building and Using Parsed Corpora.* Dordrecht: Kluwer. 2003.

Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, Magnus Åström: *The Linguistic Annotation System of the Stockholm Umeå Corpus Project*. Department of Linguistics, Umeå University.1992.

Christopher D. Manning and Hinrich Schütze: *Foundations of Statistical Natural Language Processing*. MIT Press. 1999.

Matthias T. Kromann: The Danish Dependency Treebank and the underlying linguistic theory, in Joakim Nivre and Erhard Hinrichs (eds.), *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö University Press, 2003.

Daniel Marcu, Magdalena Romera, and Estibaliz Amorrortu: Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. *The Workshop on Levels of Representation in Discourse*, pages 71-78, Edinburgh, Scotland, July 1999.

Beáta Megyesi: *Data-Driven Syntactic Analysis - Methods and Applications for Swedish.* Ph.D.Thesis. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden. 2002.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber: The Penn Discourse TreeBank. In *Proceedings of the Language Resources and Evaluation Conference*. Lisbon, Portugal. 2004.

Rosmary Stegmann, Heike Telljohann, Erhard W Hinrichs: *Stylebook for the German Treebank in VERBMOBIL*. Technical report 239. Eberhard-Karls-Universität Tübingen, 2000.

Martin Volk: *The Automatic Resolution of Prepositional Phrase - Attachment Ambiguities in German.* Habilitation thesis, University of Zurich, Faculty of Arts. Zurich: September 2001.

[AJC] Atlanta Journal-Constitution
www.ajc.com

[Anno] Annotate + NEGRA
http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

[NTN] Nordic Treebank Network
http://w3.msi.vxu.se/~nivre/research/nt.html

[Penn] The Penn Treebank Project
http://www.cis.upenn.edu/~treebank/home.html

[Penn-guide] Annotation Guidelines for the Penn Treebank
ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/

[Tiger] TIGERSearch
http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/

[Tgrep] TGrep2
http://tedlab.mit.edu/~dr/TGrep2/