Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, January 2012

Treebanks: Linking Linguistic Theory to Computational Linguistics

Anette Frank
Annie Zaenen
Erhard Hinrichs

Published by CSLI Publications

LiLT volume 7, issue 1 January 2012

Treebanks: Linking Linguistic Theory to Computational Linguistics

Anette Frank, *Heidelberg University* Annie Zaenen, *Stanford University* Erhard Hinrichs, *University of Tübingen*

Abstract

Treebanks are language resources that provide annotations at various levels of linguistic structure starting from the word level. They typically provide syntactic constituent or dependency structures for sentences, but increasingly extend to annotation beyond syntactic structure, including semantic, pragmatic and rhetorical annotation, or go beyond a single language, as in parallel treebanks.

Experience in building treebanks has shown that there is a close relation between formal linguistic theory and the design and practice of annotation. With increasing complexity of annotations, the design of annotation schemes becomes more and more theory-dependent. At the same time, linguistically motivated treebank annotations have become crucially important for the development of data-driven approaches to natural language processing and for linguistic research in general.

Treebanks therefore constitute an important link between linguistic theory and computational linguistics.

The *International Workshop on Treebanks and Linguistic Theories* provides a forum for researchers working on treebanks from both perspectives. The present volume presents the contents of the 10th edition of this workshop series, held in 2012 at the University of Heidelberg.

LiLT Volume 7, Issue 1, January 2012.
Treebanks: Linking Linguistic Theory to Computational Linguistics.

1 Ten Years of *"Treebanks and Linguistic Theories"*

"Treebanks and Linguistic Theories" was initiated as an international workshop series in 2002 in Sozopol, on the initiative of Erhard Hinrichs and Kiril Simov. Over the last ten years this workshop has successfully established itself as a forum for researchers in the field of Computational Linguistics who are experts in the design, creation and exploitation of annotated corpora and their role in linking linguistic theory and computational linguistics.

At the time this workshop was initiated, the relevance of treebanks for advancing the state of the art in Natural Language Processing was clearly acknowledged, by the impact of the first projects creating annotated corpora for English, such as the Penn Treebank, the Susanne and the Lancaster corpora, but also pioneering projects that started to build treebanks for other languages, such as German, Czech, Bulgarian, Italian or Spanish.

With the rise of more and more treebanking projects targeting novel languages and layers of annotation, there was also an increasing awareness of the relevance and impact of linguistic theory for the design of treebank annotation schemes. With more and more languages being addressed, novel linguistic phenomena needed to be accommodated, and insights from linguistic theory proved crucial for designing well-motivated, transparent, informative and consistent annotation schemes that went beyond the schemes established for English.

Since its first edition in 2002, the TLT workshop series has proven successful in attracting high-quality papers every year, presenting innovative work on various aspects relating treebanks, linguistic theories and computational linguistics.

Topics discussed comprise reports on treebank initiatives for novel languages and novel linguistic phenomena; the development of treebanks for historical data sets; the design of annotation schemes and their relation to linguistic theory; the way in which different layers of semantic information can be incorporated into treebanks, and cross-linguistic aspects in building parallel treebanks. Equally important was the discussion of issues of annotation consistency control, semi-automatic annotation techniques including the question of whether treebank annotations should be generated from an underlying treebank grammar, and finally, the exploitation of treebanks for statistically trained NLP components and for linguistic theory, supported by flexible querying and visualisation techniques.

Over the last ten years, and with increasing insights in techniques and methods for creating, maintaining and exploiting treebanks for

NLP and linguistic theory, the workshop themes have evolved considerably, both in breath and in depth. As will become apparent below, through the brief introduction to the contents of this volume.

2 Contents of this Volume

The present volume comprises 20 contributions, 13 long papers and 6 short papers that were presented in the TLT10 workshop at Heidelberg University. The program featured in addition two keynote lectures, by Victoria Rosén, University of Bergen and Eduard Hovy, University of Southern California.

Parallel Treebank Annotations and Querying Tools. The first contributions of the volume present different motivations and techniques for producing parallel syntactic annotations for treebanks that highlight the importance of linguistic theories in treebank creation and exploitation.

Wroblewska (2012)[1] presents a fully automated method that converts a constituent-based treebank for Polish to a full-fledged dependency treebank. The paper discusses in detail the characteristics of the Polish dependency types that are generated, in a controled way, by the automatic conversion process. The resulting treebank is successfully used to train a dependency parser for Polish. *Hautli et al. (2012)* propose to add an additional dependency annotation layer to a treebank of Hindi/Urdu, using annotations provided by a computational Lexical-Functional Grammar (LFG) of Urdu. The LFG f-structures are argued to offer deeper analysis of verbal structures regarding modality, tense and aspect. *Kübler et al. (2012)* introduce a treebank based on a dialogue corpus that features highly parallel syntactic annotations, including classical constituent and dependency structures, as well as annotations following Combinatorial Categorial Grammar (CCG). Comprising three types of syntactic formalisms in parallel, this treebank is particularly suited to support linguistic analysis and comparative parser evaluation.

With rising complexity of treebank annotations, offering appropriate facilities for querying and analyzing the annotations becomes an import issue, to allow for flexible analysis and exploitation of the annotations. *Søgaard and Kristiansen (2012)* present a powerful and efficient logic-based querying tool for dependency treebanks that is made publicly available for the community.

Finally, *Butler and Yoshimoto (2012)* present an approach for the automatic construction of meaning representations from syntactically

[1] Here and in the following all citations refer to papers in the present volume.

annotated treebanks. The method is flexible regarding variations in syntactic input structures and has been applied to the Penn Treebank and Brown-GENIA treebank.

Annotation of Out-of-domain and Digital Humanities Corpora Currently available statistically trained NLP tools are mostly based on treebanks built on texts from the news domain. Applying these models to out-of-domain corpora is known to be problematic, but the impact of domain changes has not yet been explored in detail. Deeper investigation of these aspects are of particular importance for the use of NLP techniques in Digital Humanities research.

The study of *Hinrichs and Zastrow (2012)* applies linguistic annotation tools for PoS tagging, lemmatization, parsing and Named Entity Recognition to a diachronically diverse corpus constructed from the German Gutenberg Project and analyzes the annotation results. *Frank et al. (2012)* focus on automatic semantic annotation of a corpus for Digital Humanities research. The paper analyzes the impact of domain adaptation strategies, and present a novel method for semantic annotation consistency control using Markov Logics. *Bertoldi and Chishman (2012)* investigate the applicability of Frame Semantics for the semantic annotation of legal texts, stressing in particular divergences between linguistic and world knowledge and the impact this has for annotating legal texts in a multi-lingual and multi-cultural context.

Analysis of Parser Errors and Impact of Annotation Schemes *Rehbein et al. (2012)* investigate the impact of PoS tagging accuracy on constitutent parsing in a learner corpus and show that certain types of errors in PoS tagging strongly affect parsing accuracy. *Husain and Agrawal (2012)* analyze parser errors in light of treebank annotation decisions. They show that careful analysis of parsing errors can help guiding treebank design decisions for improved parser performance.

Parallel Treebanks Treebank development is often addressed from a cross-linguistic perspective, with motivations ranging from theoretical interest in typological divergences to practical application perspectives, e.g. the use of parallel treebanks for Machine Translation.

Samardžić and Merlo (2012) presents a crosslinguistic study of verbal causal alternations that establishes a corpus-based measure for capturing cross-linguistic divergences in the usage of these constructions. Parallel treebanking of radically different languages – German, Spanish and Quechua – is investigated in *Gonzales et al. (2012)*. The paper investigates the effects of word alignments to German vs. Spanish as a basis for word segmentation of the agglutinative language Quechua. *Simov and Osenova (2012)* discuss the difficulty of aligning semantic

representations in parallel treebanks and establish a method for deriving semantic alignments at the phrasal level automatically from manual alignments at the word level.

Treebank Annotation Design A number of papers address diverging aspects of treebank annotation design. *Klenner et al. (2012)* propose an annotation scheme for polarity annotation for German that is directed towards compositional clause-level polarity prediction. *Mille et al. (2012)* investigate the impact of modular multi-level annotation in the framework of Meaning-Text Theory. *Muhonen and Purtonen (2012)* discuss the need of representing semantic ambiguity in treebanks, based on an empirical study on a dependency treebank of Finnish.

Bootstrapping and Error Detection for Treebank Development Bootstrapping methods for rapid development of syntactic treebanks are presented in *Seraji et al. (2012)* and *Agarwal et al. (2012)* for the construction of dependency- and HPSG-based treebanks for Persian. *Ghayoomi (2012)* investigates different methods for error detection in a manually constructed treebank for Hindi.

3 Invited Lectures

Two keynote lectures were held in the TLT10 workshop that addressed topics central for achieving enhanced deployment of treebanks.

Victoria Rosén from the University of Bergen gave a talk about *a Virtual Laboratory for Treebanking*, a goal pursued in the recently started INESS project. It aims at building a web-based infrastructure for treebanks with rich functionality, supporting dynamic parsing, disambiguation and advanced search of LFG treebanks, as well as online search and processing of dependency and constituent structure treebanks, including parallel treebanks.

Eduard Hovy from the University of Southern California talked about *Building a Large Corpus of Shallow Semantics: The OntoNotes Project*. He addressed the need for integrated semantic annotations for NLP research and how this is addressed through multi-level shallow semantic annotation of texts from of various genres and domains in multiple languages in the OntoNotes project. He highlighted the difficulties faced in semantic annotation, in particular annotation methodology and issues of annotation consistency, and how these problems were addressed in the OntoNotes project.

Acknowledgements

The TLT10 Program Committee Members, through careful review of submissions, helped to put together the scientific program for the conference and this special volume.

Johan Bos
Gosse Bouma
Koenraad De Smedt
Markus Dickinson
Stefanie Dipper
Dan Flickinger
Eva Hajičová
Erhard Hinrichs
Julia Hockenmaier
Valia Kordoni
Sandra Kübler
Detmar Meurers
Yusuke Miyao
Geertjan van Noord
Kemal Oflazer
Sebastian Padó
Marco Passarotti
Adam Przepiórkowski
Victoria Rosén
Kiril Simov
Caroline Sporleder
Manfred Stede
Martin Volk
Annie Zaenen
Heike Zinsmeister

Final decisions were taken by the TLT10 Program Committee Chairs

António Branco, Anette Frank and Kaili Müürisep.