HANDBOOK OF
NATURAL
LANGUAGE
PROCESSING
SECOND EDITION

Edited by
NITIN INDURKHYA
FRED J. DAMEREAU
8

Treebank Annotation

8.1 Introduction .............................................. 167
8.2 Corpus Annotation Types ...................................... 168
8.3 Morphosyntactic Annotation .................................... 169
8.4 Treebanks: Syntactic, Semantic, and Discourse Annotation ... 169
   • Motivation and Definition • An Example: The Penn Treebank •
   • Annotation and Linguistic Theory • Going Beyond the Surface Shape of
   the Sentence

8.5 The Process of Building Treebanks .................................. 176
8.6 Applications of Treebanks .......................................... 177
8.7 Searching Treebanks ............................................ 180
8.8 Conclusions ....................................................... 181
Acknowledgments ..................................................... 182
References ............................................................. 182

8.1 Introduction

Corpus annotation, whether lexical, morphological, syntactic, semantic, or any other, brings additional
linguistic information as an added value to a corpus. The annotation scenario might differ considerably
among corpora, but it is always based on some formalism that represents the desired level and area of
linguistic interpretation of the corpus. From the simple annotation of part-of-speech categories to the
shallow syntactic annotation to semantic role labeling to the “deep,” complex annotation of semantic and
discourse relations, there is usually some more or less sound linguistic theory behind the design of the
representation used, or at least certain principles common to several such theories.

Corpora have become popular resources for computationally minded linguists and computer science
experts developing applications in Natural Language Processing (NLP). Linguists typically look for various
occurrences of specific words or patterns to find examples or counterexamples within the theories they
build or work with, lexicographers use corpora for creating dictionary entries by looking for evidence of
use of words in various senses and contexts, computational linguists together with computer scientists
and statisticians construct language models and build part-of-speech taggers, syntactic parsers and various
semantic labelers to be used in applications, such as machine translation, information retrieval, informa-
tion extraction, question answering and summarization systems, dialogue systems and many more. Often,
annotated corpora were built by linguists who wanted to confront their theory with real-world texts.

Most of the work on annotated corpora concerns the domain of written texts, on which this chapter
is focused. However, it should be acknowledged that the growing interest in the speech community to
develop advanced models of spoken language has led to an increasing effort to process corpora that
represent the spoken form of language. This is well documented among other things by the contributions
in the special issue of the Journal Speech Communication published in 2001 (Bird and Harrington 2001), in
the agreement between annotators should be carefully watched and measured, in order to make the annotation guidelines more explicit and unambiguous.

Thanks to treebanks, NLP technologies such as automatic tagging, parsing, and other annotation of (mostly) written texts has made tremendous progress during the past 10–20 years. Part-of-speech tagging seems to be close to its current limits, reaching the level of human performance (as defined by the interannotator agreement). Parsing, "deep" parsing, semantic role labeling, machine translation, and other NLP technologies are still areas of vivid research and experimentation. It is expected that the findings accumulated during these experiments will influence future treebank annotation projects to serve better NLP technology needs. Similar influence might come from the theoretical side: new annotation schemes will then support, in the areas of syntax and semantics, (hopefully) more consistent, more adequate, and more explanatory linguistic theories than they do today.

Acknowledgments

The authors acknowledge the support of the Czech Ministry of Education (grants MSM-0021620838 and ME838), the Czech Grant Agency (project under the grant 405/09/0729), and the Grant Agency of Charles University in Prague (project GAUK 52408/2008). We are grateful to Barbora Vídová Hlaidk and Zdeněk Žabokrtský for reading and commenting upon the first draft of the chapter and for providing us with useful information and recommendations we used in the relevant places of the text, as well as to Pavel Straňák for his additions in the paragraphs on word sense disambiguation and named entities. Thanks are due to the two reviewers of the chapter Steve Bird and Adam Meyers for most helpful comments.

References