Review of *Treebanks: Building and Using Parsed Corpora*

Edited by Anne Abeillé

Philip Resnik
Department of Linguistics and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
resnik@umd.edu
Phone: (301) 405-6760
Fax: (301) 405-7104

Word count: 1777 words
Annotated corpora have been the fuel for a number of recent advances in the study of language, notably, although not exclusively, in computational linguistics. At the sentence level, corpus annotations range from shallow levels of linguistic representation, such as part-of-speech categories (Francis and Kučera, 1982; Leech, Garside, and Bryant, 1994) or named entities (Strassel, Mitchell, and Huang, 2003), through intermediate levels such as argument structure (Meyers et al., 2004; Palmer, Gildea, and Kingsbury, 2005), to deeper levels of semantic representation such as semantic roles (Baker, Fillmore, and Lowe, 1998), word senses (Landes, Leacock, and Tengi, 1998), events and temporal relations (Day et al., 2003), or language-independent meaning representations (Dorr et al., submitted). When it comes to annotating sentences with linguistic representation, the sky is the limit (Meyers, 2005).

The “sweet spot” in this range of annotations is occupied by treebanks, which is to say parsed (syntactically annotated) corpora. Unlike shallower annotations, syntactic parses capture hierarchical organization, a fundamental notion in virtually
any modern theory of sentence structure. But unlike the deeper representations, syntactic parses can be created with high levels of inter-annotator reliability. (Poor agreement between independent annotators is a sign of immaturity in the underlying theory.) With respect to natural language processing applications, the impact of treebanks during the last decade has been remarkable, more than validating Marcus et al.’s (1993) premise that “significant, rapid progress can be made ... by investigating those phenomena that occur most centrally in naturally occurring un-constrained materials and by attempting to automatically extract information about language from very large corpora”. Beyond applications, syntactically annotated corpora have also begun to play an increasingly productive role in psycholinguistics, theoretical syntax, and language pedagogy (e.g. (Corley et al., 2001; Jurafsky, 2002; Dillon, 2005; Meurers, 2005; Resnik et al., 2005)).

In *Treebanks: Building and Using Parsed Corpora*, Anne Abeillé draws together a collection of fifteen short pieces focused primarily on the issues that come up in creating treebanks, demonstrated across an impressive variety of languages, along with six chapters on how treebanks are used. Although twenty-one chapters cannot be covered in detail in a short space, I will present a brief walk through the chapters, followed by a discussion of the book as a whole.

Abeillé’s introduction offers a very concise but clearly written primer on the main issues that come up in choosing representations, annotating corpora, and us-
ing treebanks in applications, folding in the obligatory pointers to the chapters that follow.

In “The Penn Treebank: An Overview”, Anne Taylor, Mitch Marcus and Beatrice Santorini provide a short, accessible description of the widely used Penn Treebank (English), extracting and updating the seminal article by Marcus et al. (1993) (which still remains the definitive source for in-depth discussion).

In “Thoughts on Two Decades of Drawing Trees”, Geoffrey Sampson offers an engaging, personal discussion that combines elements of corpus description (SUSEANNE treebank and CHRISTINE corpus, English), linguistic analysis, and position paper. Unlike other chapters in the book, Sampson’s chapter offers a high level look at corpus linguistics as an engineering and scientific discipline, and — contrary to most current treebanking work — suggests that corpus annotation should make detail, accuracy, and explicitness a higher priority than the number of sentences annotated.

Among the next thirteen chapters, nine provide detailed discussions of treebanking projects for specific languages, following a pattern that generally includes:

- the goals and historical context of the project,

- the selection of data to annotate (usually news text),

- the annotation process (usually automatic analysis followed by human cor-
rection using project-specific tools),

- details of representation (admitting wide variety, but usually an elaboration on syntactic constituency or grammatical dependency representation, with additional features to address language-specific issues),

- tools used for manual creation and correction of annotations (again admitting very wide variety)

- a set of detailed, language-specific annotation choices that illustrate interesting and challenging aspects of the language under consideration,

- a description of the project’s status, and

- a brief evaluation and/or use case for the resulting corpus.

Within this framework, each chapter generally has at least one take-home message that distinguishes it from the others. Järvinen, in “Bank of English and Beyond” (Bank of English), emphasizes the limits of using a constraint grammar parser for the automatic analysis, notably a large amount of residual ambiguity requiring manual correction. Brants, Skut, and Uszkoreit, in “Syntactic Annotation of a German Newspaper Corpus” (NEGRA, German), describe an interactive annotation process in which the tagger and parser automatically draw attention to unreliable parts of the analysis that are likely to benefit from interactive human correction.
Böhmová, Hajič, Hajičová, and Hladká, in “The Prague Dependency Treebank” (PDT, Czech), give an overview of the PDT, emphasizing Czech’s rich morphology and free word order, as well as a “tectogrammatical” level that captures elements of meaning rather than strictly syntactic dependencies. Montemagni and numerous co-authors, in “Building the Italian Syntactic-Semantic Treebank” (ISST, Italian), also go beyond morphosyntax to include a lexico-semantic annotation based primarily on Italian WordNet. Moreno, López, and Sánchez, in “Developing a Syntactic Annotation Scheme and Tools for a Spanish Treebank” (SAC, Spanish), provide an overview of issues and problems in developing an annotation framework for Spanish. Abeillé, Clément, and Toussenel, in “Building a Treebank for French” (French Treebank, French) present a nicely detailed description of a Penn Treebank style annotation for French, producing a large, deployed treebank along with reusable tools and guidelines. Chen, Luo, Chang, Chen, Chen, Huang, and Gao, in “Sinica Treebank” (Sinica, Chinese) discuss an annotation project for Mandarin in which the syntax exploits a richly annotated lexical database, and in which they also encode thematic predicate-argument relations. Kurohashi and Nagao, in “Building a Japanese Parsed Corpus” (Kyoto University, Japanese), advocate an approach in which annotators improve the parser and grammar while correcting automatically generated annotations. Oflazer, Say, Hakkani-Tür, and Tür, in “Building a Turkish Treebank” (Turkish), discuss the challenges of annotating an
agglutinative language, using a rich set of morphological features and adopting a strategy in which syntactic dependency relations hold between word-internal inflectional groups.

The remaining four treebank project descriptions diverge from the general pattern. Wallis, in “Completing Parsed Corpora” (ICE-GB, English), advocates going beyond the conventional “longitudinal” (tree-by-tree) correction process to a “transverse” process that allows annotators to remain focused on a single phenomenon at a time. Becker, Bredenkamp, Crysmann, and Klein, in “Annotation of Error Types for German Newsgroup Corpus” (FLAG, German), annotate high “error density” texts with error types, in order to support automated testing of controlled language and grammar checking applications. Marciniak, Mykowiecka, Przepiórkowski, and Kupść, in “An HPSG-Annotated Test Suite for Polish” (BRG, Polish), are focused on a test suite, covering phenomena in written Polish. And Rocio, Alves, Lopes, Xavier, and Vicente, in “Automated Creation of a Medieval Portuguese Partial Treebank” (Universidade Nova de Lisboa, Portuguese), tackle annotation of Medieval Portuguese by using tools for Contemporary Portuguese — including a parser based on the same grammar — to bootstrap resources for that language.

In the sectio on using treebanks, six chapters cover annotation standards, evaluation, and grammar extraction. Ide and Romary, in “Encoding Syntactic Anno-
tation”, propose the XCES framework, an instantiation of the Corpus Encoding Standard, as a unified interchange format, in order to address the growing number of different and mutually incompatible annotation schemes. Carroll, Minnen, and Briscoe, in “Parser Evaluation: Using a Grammatical Relation Annotation Scheme” propose a dependency-based annotation scheme for use in evaluation, realized in a small evaluation corpus for English. Their approach can be viewed as a follow-on to Lin’s earlier proposal to move from constituencies to dependencies in parser evaluation, which he recapitulates in his chapter, “Dependency-Based Evaluation of Minipar”. Bod, in “Extracting Stochastic Grammars from Treebanks”, discusses his data-oriented parsing (DOP) framework, which uses memorized fragments of observed parse structures, exploring whether \textit{a priori} restrictions on the memorized fragments help or hurt parser performance compared to a “maximalist” approach in which all fragments are stored. (They hurt.) Neumann, in “A Uniform Method for Automatically Extracting Stochastic Lexicalized Tree Grammars from Treebanks and HPSG”, presents a technique for deriving a (i.e. probabilistic) grammar which can be applied both to treebanks and to a non-stochastic HPSG competence grammar, with the goal of achieving a grammar that combines their respective advantages. And Frank, Sadler, van Genabith, and Way, in “From Treebank Resources to LFG F-Structures”, propose techniques for (mostly) automatically adding linguistically motivated lexical-functional grammar (LFG) feature
structure annotations to phrase-structure treebanks

In attempting to form a concise, unified impression of Abeillé’s *Treebanks*, I am impressed by its scope, in terms of the range of languages discussed, the number of issues touched upon, and the range of real-world examples. At the same time, although virtually every chapter is individually interesting in some way, the book has the overall feel of a loose collection rather than a unified presentation. It may be that this perception reflects the current state of the field itself: despite some efforts in the community to provide common ground (for example, Ide and Romary’s chapter), the business of treebanking involves a wild proliferation of schemes for identifying part of speech categories, syntactic constituents, dependencies, and the like, clearly visible in the chapters’ numerous project-specific notations, tables, figures, tool descriptions, appendices. But my perception is reinforced by some of the book’s production choices, including a distracting tendency toward fuzziness and visual inconsistency in figures, as well as a somewhat eclectic “Using Treebanks” section. For example, widely influential work on treebank-based parsing in the line of Collins (1996; 1997), Charniak (1997), and others is mentioned briefly by Bod, but otherwise not represented — this editorial choice may reflect the fact that such work is covered in Manning and Schütze’s (1999) widely used textbook on statistical methods in natural language processing.

Still, as Abeillé states in her introduction, work on treebanks is evolving very
quickly, and it would certainly have been premature to attempt comprehensive, textbook-like coverage of all the issues that arise. Surveying this collection is a necessity for anyone contemplating the effort of building a treebank, and looking at the representational issues that arise is a must for anyone with a serious interest in developing linguistically informed tools for searching corpora (e.g. (Christ, 1994; Corley et al., 2001; Meurers, 2005; Resnik et al., 2005)). For computational linguists working on automatic parsing, a pass through this book should be required, in order to encourage attention to the diverse and deep problems not encountered in the Wall Street Journal portion of the Penn Treebank.

For linguists not previously familiar with work on treebanks, this may not be the easiest collection to jump into. However, a good path into the material would be to start with Sampson’s enjoyable, programmatic contribution, take a side trip to Abney’s (1996) valuable and provocative discussion of statistical methods and linguistics, return to Taylor et al.’s Penn Treebank overview chapter, and then use Abeillé’s well organized introduction as the jumping-off point into the rest of the volume. The reader following this path will be rewarded with a clear sense of the challenge and the promise of systematically applying theoretically-motivated linguistic representations to “language in the large”.

10
References


Recognition, 41st Annual Meeting of the Association for Computational Linguistics.

http://papers.ldc.upenn.edu/ACL2003/MultilingualNE.pdf.