Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

HANDBOOK OF NATURAL LANGUAGE PROCESSING

SECOND EDITION

Edited by
NITIN INDURKHYA
FRED J. DAMERAU

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-8592-1 (Hardback)

Library of Congress Cataloging-in-Publication Data

Handbook of natural language processing / [edited by] Nitin Indurkhya and Fred J. Damerau.
p. cm. -- (Chapman & Hall/CRC machine learning & pattern recognition)
Includes bibliographical references and index.
ISBN 978-1-4200-8592-1 (alk. paper)
1. Natural language processing (Computer science)--Handbooks, manuals, etc. I. Indurkhya, Nitin. II. Damerau, Frederick J. (Frederick Jacob), 1931-

QA76.9.N38H363 2010
006.3'5--dc22
2009049507

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

8

Treebank Annotation

Eva Hajičová
Charles University

Anne Abeillé
Université Paris 7 and CNRS

Jan Hajič
Charles University

Jiří Mírovský
Charles University

Zdeňka Urešová
Charles University

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, information 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 Vidová Hladká 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

Abeillé, A., Clément, L., and F. Toussenel. 2003. Building a treebank for French. In *Treebanks: Building and Using Parsed Corpora*, ed. A. Abeillé, pp. 165–188. Dordrecht, the Netherlands: Kluwer.

Adda, G., Mariani, J., Paroubek, P., Rajman, M., and J. Lecomte. 1999. L'action GRACE d'évaluation de l'assignation de parties du discours pour le français. *Langues* 2(2): 119–129.

Aduriz, I., Aranzabe, M., Arriola, J. et al. 2003. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. In *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster, U.K., eds. D. Archer, P. Rayson, A. Wilson, and T. McEnery, pp. 10–11. UCREL technical paper (16). UCREL, Lancaster University.

Arnold, J. E., Wasow, T., Losongco, A., and R. Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering, *Language* 76: 28–55.

Arstein R. and M. Poesio. 2008. Inter-coder agreement for computational LinguisticsInter-Coder agreement for computational linguistics. *Computational Linguistics* 34(4): 555–596.

Balabanova, E. and K. Ivanova. 2002. Creating a machine-readable version of Bulgarian valence dictionary (A case study of CLaRK system application). In *Proceedings of TLT 2002*, Sozopol, Bulgaria, eds. E. Hinrichs and K. Simov, pp. 1–12.

Bennett, E. M., Alpert, R., and A. C. Goldstein. 1954. Communications through limited questioning. *Public Opinion Quarterly* 18(3): 303–308.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge, U.K.: Cambridge University Press.

Bick, E. 2003. Arboretum, a Hybrid Treebank for Danish. In *Proceedings of TLT 2003*, Växjö, Sweden, eds. J. Nivre and E. Hinrich, pp. 9–20.

Bird, S., Chen, Y., Davidson, S., Lee, H., and Y. Zheng. 2006. Designing and evaluating an XPath dialect for linguistic queries. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, Atlanta, GA, pp. 52–61.