LiLT volume 18, issue 0 July 2019

Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing

Prashant Pardeshi, *NINJAL*,
Alistair Butler, *Hirosaki University*,
Stephen Horn,
Kei Yoshimoto, *Tohoku University*,
Iku Nagasaki, *Nagoya University*

Over the last few decades, corpora with comprehensive syntactic annotation, known as treebanks or parsed corpora, have been created in various formats for major languages of the world (e.g., Sampson (1995), Bies et al. (1995), Chen et al. (1999), TIGER (2003), NPCMJ (2016), etc.). As modes of accessing annotation have become more linguistically sophisticated, so these corpus resources have become more relevant for linguistics in general by providing sources of insight into factors that only become visible through analysis generalized over structures: phenomena in co-occurrence, frequency, constituency, embeddability, scope, agreement, dependency, etc. These insights are spurring new research and refinements in both corpus techniques and theoretical understanding. While much research has concentrated on challenges inherent in the creation as well as correction of annotated corpora (e.g., Dickinson and Meurers (2003), Hovy and Lavid (2010), Kulick et al. (2013), etc.), with the availability of digitized data on a large scale and the production of parsed corpora as available resources, new challenges have opened up for making use of corpus-building technologies and the resulting data in subsequent research. Examples include linking corpora to external resources like lexical databases, abstracting the contents sufficiently to be of use to non-experts, exploration of cross-linguistic patterns, etc. This special issue consists of five articles focused on applying parsed corpora research in three areas: (I) enrichment and extension of parsed corpora for dedicated uses in language processing (Pintzuk; Bin Li et al), (II) linguistic research using parsed corpora (Kubota and Kubota; Kishimoto and Pardeshi), and (III) the application of parsed corpora to language pedagogy (Wallis et al). Typologically diverse languages such as English, Japanese, and Chinese are represented, and diachronic as well as synchronic research is included. The articles in this special issue are based on presentations made at the international symposium entitled *Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing* held at the National Institute for Japanese Language and Linguistics (NINJAL) on Dec. 9-10, 2017 and organized by the collaborative research project at NINJAL entitled *Development of and Linguistic Research with a Parsed Corpus of Japanese* with which all the guest editors are associated. For this issue, each paper was lightly reviewed by two reviewers (a guest editor and an external reviewer).

References

Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. *Bracketing guidelines for Treebank II style Penn Treebank project.*. Tech. Rep. MS-CIS-95-06, LINC LAB 281, University of Pennsylvania, Computer and Information Science Department.

Chen, Keh-Jiann, Chi-Ching Luo, Zhao-Ming Gao, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, and Chu-Ren Huang. 1999. *The CKIP Chinese Treebank: Guidelines for Annotation*. ATALA Workshop IV Treebanks, Paris, June 18-19, 1999.

Dickinson, Markus and Walt Detmar Meurers. 2003. Detecting inconsistencies in treebanks. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, pages 45–56.

Hovy, Eduard and Julia Lavid. 2010. *International Journal of Translation* 22.

Kulick, Seth, Ann Bies, Justin Mott, Mohamed Maamouri, Beatrice Santorini, and Anthony Kroch. 2013. Using derivation trees for informative treebank inter-annotator agreement evaluation. NAACL 2013: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

NPCMJ. 2016. *NPCMJ - NINJAL Parsed Corpus of Modern Japanese - 2016 to present*. Available at: http://npcmj.ninjal.ac.jp/?lang=en.

Sampson, Geoffrey. 1995. *English for the Computer: The SUSANNE Corpus and analytic scheme*. Clarendon Press, Oxford.

TIGER. 2003. *TIGER Project (2003). TIGER Annotations schema*. Universites Saarlands, Universittuttgart, and Universitotsdam.