Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, October 2011

The Interaction between Linguistics and Computational Linguistics

Virtuous, Vicious or Vacuous?

Timothy Baldwin
Valia Kordoni

Published by CSLI Publications

LiLT volume 6, issue 1 October 2011

The Interaction between Linguistics and Computational Linguistics

Virtuous, Vicious or Vacuous?

Timothy Baldwin, *The University of Melbourne,* Valia Kordoni, *DFKI GmbH and Saarland University*

1 Introduction

In its infancy, computational linguistics attempted to draw heavily on theoretical linguistics, although explicit relations between computational linguistics/natural language processing and linguistics proper have never been as productive as expected. Thus, in the '60s there were some attempts to program Chomsky's transformational grammars in order to be able to parse natural language (cf., IBM in New York). Moving on to the '70s and the '80s, throughout up to the '00s, parsers have been developed on the basis of more sophisticated grammar formalisms which have been emerging one almost after the other, really, during all those decades. Those parsers have been linguistically motivated and developed in frameworks such as Lexical Functional Grammar (LFG: Bresnan and Kaplan (1982)), Functional Unification Grammar (FUG: Kay (1984)), Generalized Phrase Structure Grammar (GPSG: Gazdar et al. (1985)), Head-Driven Phrase Structure Grammar (HPSG: Pollard and Sag (1994)), Categorial Grammar (CG: Steedman (2000)), as well as Tree-Adjoining Grammar (TAG) and the so-called mildly context sensitive and model-theoretic grammar formalisms (Kallmeyer, 2010). What all those frameworks and parsers have shared with linguistics has

LiLT Volume 6, Issue 1, October 2011.
The Interaction between Linguistics and Computational Linguistics.

primarily been the belief that the determination of syntactic structure is not a self-sufficient task, but it is necessary for the determination of semantic structure.

Other instances of co-development successes between computational and theoretical linguistics over the years have come from discourse processing (Grosz and Sidner, 1986, Walker et al., 1998) and language resource development (Marcus et al., 1993, Fellbaum, 1998, Prasad et al., 2008), as well as from the significant crossover with other areas of linguistics such as lexicography (Boguraev and Briscoe, 1989, Wilks et al., 1996), psycholinguistics (Crocker, 1996, Dijkstra and de Smedt, 1996, Keller, 2001) and corpus linguistics (McEnery and Wilson, 2001, Sampson, 2001, Meyer, 2002).

Throughout the history of the field, however, there has always been a subset of computational linguistics that has largely distanced itself from theoretical linguistics, perhaps most famously in the field of machine translation (MT) where there is relatively little in most modern-day MT systems that a linguist would identify with. In the current climate of hard-core empiricism within computational linguistics it is appropriate to reflect on where we have come from and where we are headed relative to the various other fields of linguistics.

Our purpose in this special issue is partly to reflect on the status quo and, in the process, identify potential areas where greater crossover between the fields of linguistics and computational linguistics can and perhaps should occur. It is also, however, to highlight sub-areas of computational linguistics where that crossover is happening, and can be seen to have enhanced the linguistic and computational linguistic impact of the research.

This special issue stems from an EACL 2009 workshop organised by the authors, somewhat provocatively titled "Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?". It draws together extended versions of papers presented at the workshop, in addition to invited contributions from both linguists and computational linguists.

2 Overview of the Special Issue

A brief overview of the papers contained in this volume is as follows.

In an update on his seminal 1995 paper on the interaction between linguistics and statistics, Abney (2011) challenges the linguistic community to adopt more data-intensive experimentation to enable greater scientific robustness and reproducibility. He additionally provides commentary on the evolution of computational linguistics, why such a gulf

exists between computational linguistics, and what linguistics can learn from computational linguistics in terms of methodology.

Bender (2009) extends her paper from the original EACL 2009 workshop, in exploring the nature of language independence from a linguistic typological perspective, reviewing work on language independence in computational linguistics, and challenging the field to evaluate language independence systematically in a linguistically-informed manner (Bender, 2011).

Bird (2011) challenges the computational linguistics community to contribute to language preservation through the development of language resources for small languages, and proposes seven axioms for linguistic documentation of small languages. He goes on to describe the Basic Oral Language Documentation methodology as an instance of a methodology for archiving spoken language data which adheres to the proposed axioms, and further identifies opportunities for the computational linguistics community to contribute to language documentation.

Church (2011) reviews the oscillation between rationalism and empiricism in computational linguistics, and laments that the current focus on empiricism has perhaps been too successful (Church, 2011). He claims that many of the original criticisms of empiricism in computational linguistics by Pierce, Chomsky and Minsky apply to this day, and that the field ignores them at its peril. He suggests that students in computational linguistics are focusing on narrow sub-areas, and should be provided with a more rounded education, especially in linguistics.

Hajičová (2011) reflects on the relationship between linguistics and computational linguistics via her wealth of experience with structural and functional linguistics, and bidirectional crossover between linguistics and computational linguistics in the context of dependency parsing and the Prague Dependency Treebank. She concludes that linguistics has a definitive role to play in computational linguistics, e.g. in terms of understanding the structure of machine learning problems and performing feature engineering.

Johnson (2011) considers the two tasks of parsing and language acquisition. He observes that if we consider parsing as the engineering task of generating the most probable assignment of linguistic structure for a lexical input, then relatively superficial features are able to model the sorts of constraints postulated by linguistics, and robustness/open-world parsing is more important than grammaticality/closed-world parsing. He goes on to propose ways in which grammar-based parsing approaches may move towards open-world parsing. On the topic of language acquisition, he claims that statistical methods offer the means to quantitatively explore purported universals in language acquisition.

Kay (2011) draws on the foundational work of de Saussure (1915) and Zipf (1935) in reinforcing the importance of the "signifiant" (the surface realisation) and "signifié" (what it means or refers to) in any study of language, and the power laws that govern linguistic distributions. He makes the claim that machine learning-driven natural language processing ignores either the former (e.g. in unsupervised learning, where there is generally no modelling of the content of the language) or the latter, i.e. the long tail of linguistic phenomena and the impact of world knowledge on language (e.g. in statistical machine translation). He also calls upon linguistics to embrace statistics, not as a substitute for theory, but as a tool for guiding data collection and evaluating theories.

King (2011) argues that specialisation—most broadly into linguistics and computational linguistics, but more specifically into sub-fields of each—is a pragmatic necessity and indeed desirable, as it prevents the individual sub-fields from getting bogged down in contrasting/defending themselves against counter-claims about similar phenomena from other sub-fields. She makes the claim with the caveats that: (a) there are those operating across the boundaries of the sub-fields who are aware of independent work on similar topics, and cross-fertilise between the (sub-)fields; and (b) everyone remains faithful to the linguistic data.

Levin (2011) approaches the topic from the perspective of a linguist who has taught linguistics to computer scientists, and emphasises the need for computational linguistics to have an appreciation for the true variety of language. She exemplifies this via three areas: (1) linguistic diversity; (2) the long-tail of linguistic phenomena; and (3) syntactic theory as a means of error analysis.

Steedman (2011) argues that linguistics has a lot to learn from computation if it is to be relevant to computational linguistics, in terms of coverage across the full range of linguistic phenomena, computational tractability, semantic compositionality, and the true spectre of ambiguity that comes with a broad-coverage grammar. Conversely, for computational linguistics to be relevant to linguistics, the field must move away from its data-driven focus on the "short head" at the expense of the "long tail", and look to linguists for theory-driven generalisations to be able to better handle the tail.

Through this special issue, we hope to catalyse a reconsideration of the complex interaction between linguistics and computational linguistics, and position the two fields for harmonious symbiosis well into the future.

Acknowledgments

The original workshop and this paper were supported by the Erasmus Mundus European Masters Program in Language and Communication Technologies (LCT: `http://lct-master.org`), and also Australian Research Council grant no. DP0988242.

References

Abney, Steven. 1995. Statistical methods and linguistics. In K. Judith and P. Resnik, eds., *The Balancing Act*. Cambridge, USA: The MIT Press.

Abney, Steven. 2011. Data-intensive experimental linguistics. *Linguistic Issues in Language Technology* 6(2).

Bender, Emily M. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32. Athens, Greece.

Bender, Emiy M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3).

Bird, Steven. 2011. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology* 6(4).

Boguraev, Bran K. and Edward J. Briscoe, eds. 1989. *Introduction to computational lexicography for natural language processing*. London. UK: Longman Publishing Group.

Bresnan, J. and R. M. Kaplan. 1982. Introduction: Grammars as mental representations of language. In J. Bresnan, ed., *The Mental Representation of Grammatical Relations*, pages xvii–lii. Cambridge, MA: MIT Press.

Church, Kenneth. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology* 6(5).

Crocker, Matthew W. 1996. *Computational psycholinguistics: An interdisciplinary approach to the study of language*. Dordrecht, Germany: Springer.

de Saussure, Ferdinand. 1915. *Course in General Linguistics*. London, UK: Peter Owen Ltd. Translated by W. Baskin, 1959.

Dijkstra, Ton and Koenraad de Smedt, eds. 1996. *Computational Psycholinguistics*. London, UK: Taylor & Francis.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, USA: MIT Press.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Cambridge, USA: Harvard University Press.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intention and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Hajičová, Eva. 2011. Computational linguistics without linguistics? View from Prague. *Linguistic Issues in Language Technology* 6(6).

Johnson, Mark. 2011. How relevant is linguistics to computational linguistics? *Linguistic Issues in Language Technology* 6(7).

Kallmeyer, Laura. 2010. *Parsing Beyond Context-Free Grammars*. Heidelberg, Germany: Springer.

Kay, Martin. 1984. Functional unification grammar: A formalism for machine translation. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 75–78. Stanford, California, USA: Association for Computational Linguistics.

Kay, Martin. 2011. Zipf's Law and *L'Arbitraire du Signe*. *Linguistic Issues in Language Technology* 6(8).

Keller, Frank. 2001. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, The University of Edinburgh.

King, Tracy Holloway. 2011. (Xx*-)Linguistics: Because we love language. *Linguistic Issues in Language Technology* 6(9).

Levin, Lori. 2011. Variety, idiosyncracy, and complexity in human language and language technologies. *Linguistic Issues in Language Technology* 6(10).

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics* 19(2):313–330.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: an Introduction*. Edinburgh, UK: Edinburgh Univ Press, 2nd edn.

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge, UK: Cambridge University Press.

Pollard, Carl and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago, USA: The University of Chicago Press.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968.

Sampson, Geoffrey. 2001. *Empirical Linguistics*. New York, USA: Continuum.

Steedman, Mark. 2000. *The Syntactic Process*. Cambridge, USA: MIT Press.

Steedman, Mark. 2011. Romantics and revolutionaries: What theoretical and computational linguists need to know about each other. *Linguistic Issues in Language Technology* 6(11).

Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince, eds. 1998. *Centering Theory in Discourse*. Clarendon Press: Oxford.

Wilks, Yorick, Brian M. Slator, and Louise M. Guthrie. 1996. *Electric Words: Dictionaries, Computers and Meanings*. Cambridge, USA: MIT Press.

Zipf, George K. 1935. *The Psychobiology of Language*. Boston, USA: Houghton-Mifflin.