Last Words

What Science Underlies Natural Language Engineering?

Shuly Wintner*
University of Haifa

One of the most thought-provoking proposals I have heard recently came from Lori Levin during the discussion that concluded the EACL 2009 *Workshop on the Interaction between Linguistics and Computational Linguistics*. Lori proposed that we should form an ACL Special Interest Group on *Linguistics*. At first blush, I found the idea weird: Isn't it a little like the American Academy of Pediatrics forming a SIG on Medicine (or on Children)? Second thoughts, however, revealed the appropriateness of the idea: In essence, linguistics is altogether missing in contemporary natural language engineering research. In the following pages I want to call for the return of linguistics to computational linguistics.

The last two decades were marked by a complete paradigm shift in computational linguistics. Frustrated by the inability of applications based on explicit linguistic knowledge to scale up to real-world needs, and, perhaps more deeply, frustrated with the dominating theories in formal linguistics, we looked instead to corpora that reflect language use as our sources of (implicit) knowledge. With the shift in methodology came a subtle change in the goals of our entire enterprise. Two decades ago, a computational linguist could be interested in developing NLP applications; or in formalizing (and reasoning about) linguistic processes. These days, it is the former only. A superficial look at the papers presented in our main conferences reveals that the vast majority of them are engineering papers, discussing engineering solutions to practical problems. Virtually none addresses fundamental issues in linguistics.

There's nothing wrong with engineering work, of course. Every school of technology has departments of engineering in areas as diverse as Chemical Engineering, Mechanical Engineering, Aeronautical Engineering, or Biomedical Engineering; there's no reason why there shouldn't also be a discipline of Natural Language Engineering. But in the more established disciplines, engineering departments conduct research that is informed by some well-defined branch of science. Chemical engineers study chemistry; electrical engineers study physics; aeronautical engineers study dynamics; and biomedical engineers study biology, physiology, medical sciences, and so on.

The success of engineering is also in part due to the choice of the "right" mathematics. The theoretical development of several scientific areas, notably physics, went alongside mathematical developments. Physics could not have accounted for natural phenomena without such mathematical infrastructure. For example, the development of (partial) differential equations went hand in hand with some of the greatest achievement in physics, and this branch of mathematics later turned out to be applicable also to chemistry, electrical engineering, and economics, among many other scientific fields.

* Department of Computer Science, University of Haifa, 31905 Haifa, Israel. E-mail: shuly@cs.haifa.ac.il.

What branch of science, then, underlies Natural Language Engineering? What is the theoretical infrastructure on which we build our applications? And what kind of mathematics is necessary for reasoning about human languages?

Consider some of the greatest achievements of natural language engineering since the data-oriented revolution. The Penn Treebank, for example, whose annotation has been used for training numerous POS taggers and parsers since its first release in 1992: What theory underlies its annotation? In what sense is this annotation "correct"? Could any other annotation scheme be just as good? What criteria do we have for evaluating the quality of this resource? And what branch of science should such criteria be embedded in?

Or take machine translation, the holy grail of natural language processing for half a century. We now have statistical machine translation systems that perform well enough to be usable for a variety of applications, and Google provides free machine translation services between any pair of over 40 languages, so one can translate automatically between Albanian and Vietnamese. This is probably the greatest achievement of our field; what branch of science is it based on? What theory underlies it?

I could go on and on. Word sense disambiguation, stochastic parsing, text categorization, question answering, semantic role labeling, speech recognition, ontology development, whatever your favorite application is: What branch of science underlies it? What are its theoretical underpinnings?

In the Old Days this used to be linguistics. Morphological analyzers reflected the accumulated wisdom of researchers in morphology and phonology. The first parsing algorithms were informed by syntactic theory. Dialog systems were based on research in semantics and discourse theory. Why is this no longer the case?

One reason, obviously, is that applications that were based on explicit linguistic knowledge didn't scale up well. A more subtle reason has to do with the way science is funded: Funding agencies (mainly in the U.S.) are motivated by short-term practical goals, and are less patient with long-term, infrastructural basic research. Other areas of computer science shift from foundational, theory-based research to engineering application for the same reason.

But there is a deeper reason. Linguistics, as a discipline, went astray: It focused mainly on syntax (and predominantly on English); and its theory became so obscure, so baroque, and so self-centered, that it became virtually impenetrable to researchers from other disciplines. To use the terminology of Evans and Levinson (in press), "the relevant literature is forbiddingly opaque to outsiders"; or, in the words of Tomasello (1995, page 136), linguistic theories are "described in linguistically specific terms such that it is very difficult to relate them to cognition in other psychological domains." Or to computational implementation, for that matter.

So we were frustrated with linguistics, and in our frustration we totally abandoned it, and were left with statistics and probability theory. But surely computational linguistics cannot be a branch of applied statistics. For if this were the case, nothing could distinguish natural languages from other, *non*-linguistic, string manipulation systems, such as DNA sequences or music score sheets or transcripts of chess games. Surely there's something *unique* to the strings that our systems manipulate, something that can be theorized about and can be scientifically investigated. What makes our systems special is the fact that they manipulate natural languages, and the only scientific field that can inform our work is linguistics.

And the truth is that there's much new in the world of linguistics, much that should interest us computational linguists. The tight grip of generative syntax on the world of theoretical linguistics has long been released, and there are several excellent

research directions that could greatly benefit from a more formal, mathematical, and computational investigation. Let me illustrate using just a few examples.

The most prominent example is psycholinguistics. Several researchers with a genuine interest in language, but with training in the cognitive sciences, address linguistics in a more general, cognitive context, and bring to the investigation promising scientific methodologies. Consider CHILDES (MacWhinney 2000), a vast computational corpus transcribing linguistic interactions between children and their caretakers, in over 25 languages, collected over more than 20 years by dozens of researchers. These data are for the most part annotated morphologically (and, to some extent, also syntactically). They have been used in over 1,500 scientific papers for investigating and evaluating diverse issues in language development. They provide an invaluable resource for computationally driven research in linguistics: psycholinguistics, theoretical linguistics, cognitive linguistics and, yes, also computational linguistics.

Psycholinguistics provides not only resources and research methodologies, but also theories that we, as computational linguists, should be able to inspect, evaluate, and elaborate on. Researchers including MacWhinney (1987, 1998, 1999, 2004a, 2004b), Tomasello (1998, 2003, 2006), or Bybee (2001, 2006, 2007), to mention just a few, produce exciting theories that are backed up by experimentation. Interestingly enough, these linguistic theories hint at the kind of mathematics that we need to develop in order to understand and reason about natural languages: They emphasize language use over abstract "competence" and direct us from formal grammars and logic to statistics and probability theory. Most importantly, such theories are in principle falsifiable (Tomasello 2004); can we improve them using our own, mathematical and computational, methodologies? Can we use them to build better systems?

Exciting linguistic research, which is grounded in more general, philosophical, biological, cognitive, and computational insights, is also performed in language evolution (Christiansen and Kirby 2003), in historical linguistics (Warnow 1997; Nakhleh, Ringe, and Warnow 2005), and also in the more traditionally central areas of morphology and syntax (Goldberg 1995; Prince and Smolensky 1997; Pullum 2007). We must be aware of this work.

But we can go much further. Not only should we be more aware of linguistic research that can improve our engineering work, we should also be directly involved in such research. Our formal mathematical and computational training is invaluable for research in the life sciences and the humanities, and can shed new light on phenomena that traditional approaches fail to account for. We can bring refreshing insights and new points of view to all branches of linguistics. Computational linguistics can, in essence, be a sub-field of linguistics.

In fact, some first attempts in this direction have already seen their way into our main conferences. For example, Ellison and Kirby (2006, page 273) propose a computational method for constructing genetic language taxonomies, which they claim "coheres better with current thinking in linguistics and psycholinguistics." Daumé III and Campbell (2007) use a computational methodology, applied to the World Atlas of Language Structures (a large database of various properties of over 2,000 languages), to discover linguistic universals, in the form of typological implications. In both of these examples, novel computational techniques are used to assist in exactly the type of research that traditional linguists have always been interested in, and both cases are success stories.

Such examples, unfortunately, are still few and far between. But they demonstrate what computational linguistics can achieve when it is backed up and informed by linguistic theory. Whether such theoretical research is guaranteed to improve engineering

projects I cannot tell; but in all areas of the natural sciences this has been the case. Language engineering should not be different.

More importantly, our community is able to be a major force behind contemporary research in linguistics. Many of us were drawn to this field by our love for language; we can now follow our hearts and go back to exploring language in all its beauty, rather than (okay, in addition to) harnessing it to our practical needs. Let us be part of linguistics. Or do some of us really have to settle for being refugees in an ACL SIG?

Acknowledgments

I wish to thank Robert Dale for inviting me to write this column and for his help with its preparation. I am grateful to Nissim Francez and Nurit Melnik for commenting on an earlier draft of this piece. The views expressed here, including all errors and misconceptions, are of course my own.

References

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge University Press, Cambridge.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford University Press, Oxford.

Christiansen, Morten H. and Simon Kirby, editors. 2003. *Language Evolution*. Studies in the Evolution of Language. Oxford University Press, Oxford.

Daumé III, Hal and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Conference of the Association for Computational Linguistics (ACL)*, pages 65–72, Prague.

Ellison, T. Mark and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 273–280, Morristown, NJ.

Evans, Nicholas and Stephen Levinson. In press. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*.

Goldberg, Adele. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL.

MacWhinney, Brian, editor. 1987. *Mechanisms of Language Acquisition*. Lawrence Erlbaum Associates, Hillsdale, NJ.

MacWhinney, Brian. 1998. Models of the emergence of language. *Annual Review of Psychology*, 49:199–227.

MacWhinney, Brian, editor. 1999. *The Emergence of Language*. Carnegie Mellon Symposia on Cognition. Lawrence Erlbaum Associates, Mahwah, NJ.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition.

MacWhinney, Brian. 2004a. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, 31:883–914.

MacWhinney, Brian. 2004b. A unified model of language acquisition. In J. Kroll and A. De Groot, editors, *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press, Oxford, pages 49–67.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420.

Prince, Alan and Paul Smolensky. 1997. Optimality: From neural networks to universal grammar. *Science*, 275:1604–1610.

Pullum, Geoffrey K. 2007. The evolution of model-theoretic frameworks in linguistics. In *Proceedings of the ESSLLI 2007 Workshop on Model-Theoretic Syntax*, pages 1–10, Dublin.

Tomasello, Michael. 1995. Language is not an instinct. *Cognitive Development*, 10:131–156.

Tomasello, Michael. 1998. The return of constructions. *Journal of Child Language*, 25:431–442.

Tomasello, Michael. 2003. *Constructing a Language*. Harvard University Press, Cambridge and London.

Tomasello, Michael. 2004. What kind of evidence could refute the UG hypothesis? *Studies in Language*, 28(3):642–645.

Tomasello, Michael. 2006. Acquiring linguistic constructions. In D. Kuhn and R. Siegler, editors, *Handbook of Child Psychology*. Wiley, NY, pages 255–298.

Warnow, Tandy. 1997. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences*, 94(13):6585–6590.