Linguistic Issues in Language Technology – **LiLT**
Submitted, October 2011

(Xx*-)Linguistics: Because We Love Language

Tracy Holloway King

Published by CSLI Publications

LiLT volume 6, issue 9 October 2011

(Xx*-)Linguistics: Because We Love Language

Tracy Holloway King, *Microsoft Corp.*

1 A focus on language

Having been trained as a theoretical linguist[1] and become a computational linguist who focuses on grammar engineering, I wanted to discuss what binds the two fields and why the dichotomies between the fields can be viewed as strengths. Theoretical and computational linguistics reflect the same motivations, but their realization is strikingly different. I am reminded of when I was asked to review Željko Bošković's theoretical linguistics book on clitics and the phonology-syntax interface (Bošković, 2001) for *Language* (King, 2004): I agreed whole-heartedly with his fundamental theoretical claim that clitic placement is the result of the interaction of prosodic and syntactic factors; however, I disagreed with almost every detail of his analysis. The relation between theoretical and computational linguistics often has the same feel: even when the fundamental ideas are agreed upon, the two fields have vastly different approaches to them.

What binds theoretical and computational linguists is a love of language and a desire to discover the patterns that underly its complexity. What we believe these patterns are varies significantly from linguist to linguist and school of thought to school of thought, but fundamentally

[1] In this article I use 'theoretical linguistics' to refer to the linguistics that contrasts with fields such as computational linguistics, sociolinguistics, and psycolinguistics. The term is solely for convenience: All of these fields involve theories.

LiLT Volume 6, Issue 9, October 2011.
(Xx-)Linguistics: Because We Love Language.*

we all believe that there is method to the madness and that by careful observation we will unearth the patterns.

Some theoretical linguists spend a lifetime describing a single language, documenting its surface realization and recording its underlying structure for future researchers, teachers, and language learners. For more studied languages, a linguist might focus on a set of constructions, building up a set of analyses that work together to account for the patterns of that language. Other linguists are fascinated by a given phenomenon, such as causative formation, and look at the phenomenon across languages to see the range of data that any theory must account for. In each of these approaches, the linguist believes that what they discover will extend further: to other languages, to other phenomena within the language. And, if the analyses do not extend correctly, this is taken as a signal that some generalization was missed, that some underlying pattern should have been captured but was not.

These same approaches exist in computational linguistics. Some computational linguists focus on a given language[2] with the firm, if sometimes mistaken, belief that the techniques they propose will extend to other languages (see Bender (this volume) and Abney (this volume)). As with theoretical linguistics, it is assumed that adjustments and parameterization will be necessary in order to account for the typological diversity of language, but that a pattern used for one language can be exploited in similar languages and that as we analyze typologically diverse languages, we will discover a finite and understandable set of analyses that will cover natural language. Some computational linguists embrace a particular phenomena or task and apply it to a broad set of languages to understand the diversity from the outset. This approach is often reflected in shared tasks such as those for part-of-speech tagging, noun phrase chunking, and dependency parsing ((Cardie et al., 2002) and any other CoNLL proceedings).

One thread of computational linguistic research that often horrifies theoretical linguists, and many computational linguists, is one where as little overt linguistic theory as possible is used to accomplish a task. For example, how good a machine translation system can be built using a corpus of aligned sentence pairs? How little training data is needed to produce a part of speech tagger? Can synonym sets and ontologies be learned from unanalyzed text? Although I personally prefer more overt linguistics in the work I do, such approaches have an important role to play in our understanding of language. They define a limit that

[2] And, in an even greater bias than in theoretical linguistics, this language is often English or perhaps Chinese (see Bender (2009) and Bender (this volume)).

such approaches cannot exceed and which the addition of linguistic information can be measured against.[3] For example, when part-of-speech tagging highly inflected languages like Dravidian, the use of morphological information improves performance above the baseline because data sparsity reduces the effectiveness of less linguistically-rich approaches (Sangal et al., 2007). In addition, such systems often use more linguistic knowledge than meets the eye: feature design is key in such systems and these features often reflect linguistic insights; annotated data may be used to train parts of the system and the annotations are driven by linguistic analysis.

2 It's all about the data

Both theoretical and computational linguistic research is fundamentally about language data. Researchers take a subsection of the data (no one works on analyzing all languages at all levels for all applications), analyze it, and then extend their analysis to the next set of data, be they new constructions, new languages, new corpora, or new levels of analysis.

The advent of large-scale searchable corpora, including the web, and of annotated corpora[4] have allowed linguists to use data in ways that were not possible a few years ago. For languages with a significant amount of text on the web, it is now possible to search for constructions to determine their natural distribution. For example, in Dalrymple et al. (1998), we needed to test aspects of our analysis of Russian comitatives. We were worried that artificially created examples would bias speaker judgements. By choosing relatively common words with the features we were interested in, we were able to obtain enough naturally occurring examples to refine our analysis. Lesser studied and endangered languages also benefit from the increased accessability to corpora (Abney, this volume), even if they are only available in archives. Having them electronically accessible ensures that important data will not be lost and allows those studying the language to examine large amounts of data to discover patterns and form analyses.

Another aspect of having more data readily available is that the pos-

[3] And it is impressive what the informed use of n-grams can accomplish in a system. However, see Church (this volume) on how computational linguistics is reaching the limits of what can be done with approximations.

[4] The process of producing annotated corpora, especially gold standard corpora such as those provided by the Linguistic Data Consortium, can provide a detailed view into the structure of a language. These gold standards generally pick a corpus, e.g. a year's worth of newspaper articles, and produce the relevant annotation, e.g. syntactic trees, for every sentence in the corpus. This approach forces a breadth of coverage that highlights areas where the theory does not provide an analysis.

sible extensions of an analysis to new languages or new constructions can more easily be tested. Linguists, whether computational or theoretical, propose theories, techniques, and analyses with the assumption that their proposal will extend to other languages or at least to other constructions within the language they are working on. Such proposals almost always require changes in order to adequately, much less elegantly, account for new data. The continual evolution of analyses is part of linguistic research, and having a broader set of data available as an integral part of research will aid linguists in all fields.

As more data becomes available, through the web, through archives, and through initiatives to encourage publishing data (Bender et al., 2009), linguists in all fields will use increasingly data-driven methods and will use subsets of the same data. Having different linguists with different interests and objectives examining the same data will provide the diversity of perspectives and efforts that is necessary to increase our understanding of language and its structure. Over time, these analyses will develop common aspects and hence provide evidence for correctness.

3 Is specialization truly a problem?

The different approaches to linguistics and language and the concern that linguists in different fields are not communicating is one that is frequently raised and that is not restricted to theoretical and computational linguistics. This concern is raised for all the “hyphenated” fields, and also within subfields of traditional theoretical linguistics, e.g. phonology, morphology, syntax and semantics. I would like to suggest that this is not a problem and in fact is a strength. The subfields have different views on language because of their interests: what is a prominent issue in child language acquisition may be relatively unimportant in computational linguistics; what is a controversial issue in syntax may have little impact on phonology. These different concerns allow the field to move forward on all fronts.

The issue that then arises is what happens when different theories and analyses are developed for the same phenomenon by different fields. This occurs constantly within a given field when researchers differ on their analysis, e.g. case marking (see (Butt, 2006) for an overview of case data and theory). Similarly, the same basic analyses or patterns can be independently discovered and developed across fields. This independent discovery of language structure as it relates to fields with different immediate tasks can be construed as evidence for correctness. An example of this would be parts of speech. Most fields of linguistics make use of basic classes of parts of speech and generaly agree to the classes. However, some research focuses on this issue, highlighting linguistically complex cases including the universality of the categories used. Theoretical morphology and computational linguistics, including large-scale multi-lingual annotation tasks, bring these part-of-speech issues to the fore (see Hajičová (this volume) for other examples of how large-scale gold standard creation tests and develops theoretical analyses).

In order to see the contradictions and convergences across fields, there must be researchers who work in both fields or at least follow the research in the other field. This situation already arises naturally between theoretical and computational linguistics. Both fields have researchers who were originally trained in the other field. This is particularly true in certain subfields of computational linguistics such as grammar engineering where many of the researchers came from backgrounds in theoretical syntax. On the flip side, computer scientists come to theoretical linguistics via computational linguistics. Attendance at conferences and workshops also provides cross-fertilization. Theories like LFG and HPSG, which have strong theoretical and computational researchers, host conferences where both fields are represented.[5] The larger conferences seem to be more divided.[6] This reflects a greater reality in specialization: at a given large conference, most people will attend talks within their speciality, e.g. morphology, and only occassionally attend a session on something new, e.g. semantics, to learn more about it. Those researchers who bridge the (sub)fields help ensure that, although specialization is a necessity for exploring theories and phenomena in depth, computational and theoretical linguists can profit from advances in both fields to test and refine their hypotheses. Steedman (this volume) argues that computational linguists can take advantage of theoretical linguistics to capture the long tail of phenomena which are difficult to learn from corpora, even extremely large

[5] For example, at the 2009 LFG conference, 13 of the 41 papers were computational, or had a significant computational component.

[6] Efforts have been made to change this, with limited success. On the theoretical side, the 2009 LSA meeting had a plenary symposium and an invited session on computational linguistics, but no session or posters on computational linguistics. The 2008 LSA meeting had no computational linguistics. The 2007 meeting had a tutorial on the use of databases in field linguistics (Good and Johnson, 2007) and three posters on computational linguistics. On the computational side, the 2009 EACL meeting had a workshop on the interaction between linguistics and computational linguistics (Baldwin and Kordoni, 2009), and the call for papers and the review form states at the top *EACL 2009 invites papers on all aspects of computational linguistics, theoretical and empirical, linguistic and computational, fundamental and applied.*

corpora. Similarly, Church (this volume) argues that the “low hanging fruit” captured by approximation techniques is almost exhausted and that progress in computational linguistics depends on integrating richer linguistic structures.

4 Hybridization as a common trend

Computational linguistics is seeing an increase in hybrid systems which use both rule-based and stochastic components. Such systems are also appearing within theoretical linguistics, e.g. stochastic Optimality Theory analyses for phonology and syntax (Bresnan, 2007), and within psycholinguistcs (Jaeger and Snider, 2008). For example, stochastic Optimality Theory captures variation in realization within a given speaker or across speakers. The more traditional theoretical OT constraints are not only ranked but also associated with probabilities which allow the rankings to appear to shift relative to one another.

There are many ways to hybridize a system, and which type of hybridization to use depends both on the goal of the system (e.g. part-of-speech tagging, syntactic analysis, spoken dialog production) and the phenomena of the languages being analyzed (e.g. rich morphology, fixed word order, discontinuous constituents). Some systems capture a few absolutes with rules while otherwise being a predominantly stochastic system. For example, Guo (2009)’s broad-coverage LFG grammar of Chinese is a predominantly stochastic system with a probalistic CFG which uses a rule-based component to resolve empty pronouns. In contrast, the ParGram (Butt et al., 2002) English grammar is predominantly rule-based, but used stochastic techniques to rank the parsers so that applications can use the n-best parses (Riezler et al., 2002). Some research focuses directly on optimizing hybridization. For example, Liao et al. (2009) discuss an experiment in which rule-based and stochastic components were interleaved in various ways to determine which combination yielded the best results for name tagging, with the optimal configuration using both types of components.

The DCU cross-linguistic grammar development effort (Cahill et al., 2005) brings up an interesting point: the output that the DCU grammars produce is the result of decades of theoretical linguistic research in LFG theory (Bresnan, 1982, Dalrymple, 2001), but the way in which these representations are derived draws heavily on computational linguistics and does not resemble the grammar rules found in theoretical LFG papers. And even those rules which do bear a resemblance to the theoretical LFG rules are not applied in a way that most theoretical syntacticians would apply them. So, is this a hybridized system? Regardless of the answer, it is a line of computational research that is majorly informed by theoretical linguistics: theoretical linguistics provides the representation, while computational linguistics provides the mechanisms by which those representations are realised.

The rise of hybrid analyses of language across fields of linguistics reflects how different fields can come to the same conclusions about language: that language has both stochastic and rule-driven aspects to it (see Kay (this volume) for more discussion). Focusing on one aspect or the other can drive research on particular aspects of language, but having some research which focuses on the combination can ultimately tell us more. The "hyphenated" fields of linguistics, including computational linguistics, have paid greater attention to how stochastic effects influence the use of language. Although there are researchers in theoretical linguistics who focus on these effects, it is not the norm. I would argue that this is not a problem for the field. Theoretical linguistic research focuses on providing hypotheses and analyses; stochastic effects are seen over these. As more cases where stochastic effects predominate are found, more work will be done within theoretical linguistics to account for these. In the mean time, there is still ample data to be explored in a non-stochastic context for those who prefer to pursue that avenue of research.

5 Specialist or generalist: just stay true to the language

Overall, although there is not the intense overlap between theoretical and computational linguistics that many seem to wish, I do not see this as a problem from a practical perspective. Everyone is working towards a common goal to understand language and to unearth its structure. By approaching the problem from different angles and with different immediate goals, we gain a broader perspective on the problem, and language is complex enough to need this broad perspective. As with all fields and as with the (Xx*-)linguistic fields, there are some researchers who more naturally reach outside their immediate domain and who will see where there are convergences and contradictions. Forcing everyone into this role would slow the growth of the field as a whole given the vast amount of work still before us. Over their careers, linguists can move from one field to another, just as they can explore new areas of research within a field. So, my advice is to pursue what interests you, and to continue to share your ideas and findings so that others can use them in their research.

Acknowledgments

I would like to thank the editors of this volume and the two anonymous reviewers, one of whom is the only person who has ever accused me of being Pangolossian.

References

Baldwin, Timothy and Valia Kordoni, eds. 2009. *The Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*. Associal of Computational Linguists.

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: Association for Computational Linguistics.

Bender, Emily M., Scott Farrar, Jeff Good, and Laura Welcher. 2009. Cyberling2009. Workshop at the LSA Summer Institute, Berkeley, CA; wiki and blog available at http://cyberling.elanguage.net/.

Bošković, Željko. 2001. *On the Nature of the Syntax-Phonology Interface: Cliticization and Related Phenomena*. Elsevier.

Bresnan, Joan, ed. 1982. *The Mental Representation of Grammatical Relations*. The MIT Press.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld, eds., *Roots: Linguistics in Search of Its Evidential Base*, pages 77–96. Mouton de Gruyter.

Butt, Miriam. 2006. *Theories of Case*. Cambridge University Press.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Cahill, Aoife, Martin Forst, Michael Burke, Mairéad McCarthy, Ruth O'Donovan, Christian Rohrer, Josef van Genabith, and Andy Way. 2005. Treebank-based acquisition of multilingual unification grammar resources. *Journal of Research on Language and Computation* pages 247–279.

Cardie, Claire, Walter Daelemans, Claire Nédellec, and Erik Tjong Kim Sang, eds. 2002. *Proceedings of CoNLL-2000 and LLL-2000*.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. Syntax and Semantics. Academic Press.

Dalrymple, Mary, Irene Hayrapetian, and Tracy Holloway King. 1998. The semantics of the Russian comitative construction. *Natural Language and Linguistic Theory* 16:597–631.

Good, Jeff and Heidi Johnson. 2007. A field linguist's guide to making long-lasting texts and databases. Tutorial at the 2007 meeting of the Linguistic Society of America Annual Meeting.

Guo, Yuqing. 2009. *Treebank-based Acquisition of Chinese LFG Resources for Parsing and Generation*. Ph.D. thesis, Dublin City University.

Jaeger, T. Florian and Neil Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, pages 827–812.

King, Tracy Holloway. 2004. Review of "On the nature of the syntax-phonology interface: Cliticization and related phenomena". *Language* pages 843–846.

Liao, Wenhui, Marc Light, and Sriharsha Veeramachaneni. 2009. Integrating high precision rules with statistical sequence classifiers for accuracy and speed. In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009)*, pages 74–77. Association for Computational Linguistics.

Riezler, Stefan, Tracy Holloway King, Ron Kaplan, Dick Crouch, John T. Maxwell III, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Sangal, Rajeev, Sushme Bendre, Dipti Misra Sharma, and Prashanth Reddy Mannem, eds. 2007. *Proceedings of the IJCAI-2007 Workshop on Shallow Parsing for South Asian Languages*.