Appeared in *Cognitive Science* 26(3), May-June 2002. [May have minor differences from published version.]

Introduction to the Special Section on Linguistically Apt Statistical Methods

Jason Eisner
Department of Computer Science
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218-2691 U.S.A.
Phone: +1 410 516 8775 / Fax: +1 410 516 6134
Email: jason@cs.jhu.edu

In 1994—about six years after it was first infiltrated by statistical methods—the Association for Computational Linguistics hosted a workshop called "The Balancing Act: Combining Symbolic and Statistical Approaches to Language" (Klavans & Resnik, 1996). The workshop argued that linguistics and statistics were not fundamentally at odds, even though the recent well-known statistical techniques for part-of-speech disambiguation (Church, 1988; DeRose, 1988) had, like their predecessors in speech recognition, flouted Chomsky's (1957) warnings that Markov or *n*-gram models were inadequate to model language. The success of these Markovian techniques had merely established that empirically estimated probabilities could be rather effective even with an *impoverished* theory of linguistic structure. As an engineering matter, the workshop argued, it was wise to incorporate probabilities or other numbers into *any* linguistic approach.

Several years later, it seems worth taking another snapshot from this perspective. It is fair to say that a greater proportion of hybrid approaches to language now are cleanly structured rather than cobbled together, and that the benefits to both sides of such approaches are better understood. The prevalent methodology is to design the form of one's statistical model so that it is capable of expressing the kinds of linguistic generalizations that one cares about, and then to set the free parameters of this model so that its predicted behavior roughly matches the observed behavior of some training data.

The reason that one augments a symbolic generative grammar with probabilities is to make it more robust to noise and ambiguity.[1] After all, statistics is the art of plausibly reconstructing the unknown, which is exactly what language comprehension and learning require.

Conversely, one constrains a probability model with grammar to make it more robust to poverty of the stimulus. After all, from sparse data a statistician cannot hope to estimate a separate probability for every string of the language. All that is practical is to estimate a moderate set of parameters that encode high-level properties from which the behavior of the entire language emerges.

Carrying out this program is not trivial in practice. Patterning a statistical model after a linguistic theory may require some rethinking of the theory, especially if the model is to be elegant and computationally tractable. And there is more than one way to do it: the first few tries at adding linguistic sophistication often hurt a system's accuracy rather than helping it. More complex linguistic representations also call for more complex, slower, and/or more approximate algorithms to estimate the parameters of the statistical model.

Nonetheless, the paradigm has enabled progress in many areas of linguistics, speech processing, and natural language processing. The present special section of brief reports spans diverse interests:

- Johnson and Riezler show how a model of the relative probabilities of parse trees can be made sensitive to any linguistic feature one might care to specify. They report that this approach can be applied to the tricky case of Lexical-Functional Grammar (LFG).

- Eisner explains how to attach probabilities to lexicalized grammars, including the lexical redundancy rules that express transformational generalizations in the grammar. The model is designed so that learners are naturally inclined to discover and use such generalizations.

- Light and Greiff review several published techniques for discovering lexical selectional preferences. In these techniques, the models are constrained not by just by the abstract theory of a taxonomy of meaning, but by the particular taxonomy of the WordNet lexical database.

- Nock and Young report on speech modeling techniques inspired by the fact that speech is produced not by a monolithic mouth, but by a system of articulators (tongue root, lips, etc.) that act somewhat independently of one another.

Several popular statistical techniques are used repeatedly across these papers. The first two papers use log-linear (or

[1] Students should read the charming paper by Abney (1996), in a book based on the Balancing Act workshop, in which he convincingly argues that ambiguity is pervasive and that human knowledge of language includes knowledge of the probable as well as the possible.

maximum-entropy) models, the last three build on Hidden Markov Models, and the middle two are inspired by Bayesian networks (or directed graphical models). All four invoke the Expectation Maximization algorithm for parameter estimation. And as is customary in the field, all of these researchers evaluate the performance of their methods on real data, such as newspaper text.

This research paradigm has rapidly taken over most of computational linguistics, which a decade ago was mainly concerned with pure symbolic manipulation. How does it relate to other traditions of language research?

The statistical tradition places less emphasis on the surprising boundary cases that are dear to pure linguists. Such cases can of course demonstrate that a formal framework needs to be extended or revised. But a formal framework also needs to be rich enough to describe the vagaries of common cases, to which speakers and hearers devote considerable resources. Even in core phenomena such as simple sentences, humans display a great deal of "soft" knowledge about selectional preferences, stylistics, prosodic contours, and so forth. This kind of knowledge is crucial to working systems for language and speech processing. As an engineering matter, it is apparently more important to tune the common cases than to recognize the existence or non-existence of the rare cases, which is why the community tends to start with simple models and add sophistication gradually.

Psycholinguistic research does recognize the importance of frequency effects in language processing and learning, but it too tends to differ in style from computational approaches. Psycholinguists have demonstrated, by means of ingenious experimental design and significance tests, that humans are indeed sensitive to this or that frequency variable. But computational linguists often wish their computers to actually mimic humans, so they need to create complete models that happen to exhibit these sensitivities. It is one thing to say that several variables separately affect human performance, and another to combine those variables into a quantified prediction of what the human will do. (The same might be said of modeling the stock market.)

Within the cognitive sciences, then, it is currently the computer scientists who are most committed to thinking about human language in terms of statistical models. Is this the natural state of affairs? Ultimately, the computer scientists are designing statistical models to do a better job of modeling either naturally occurring or elicited data. But that is also the job of psychologists, linguists, and perhaps language users themselves. As the available models go from being cheap and effective to being expensive but psychologically or linguistically plausible—the goal of the work outlined in these brief reports—perhaps the paradigm shift will spread to the other cognitive sciences of language.

References

Abney, S. (1996). Statistical methods and linguistics. In J. L. Klavans & P. Resnik (Eds.), *The balancing act: Combining symbolic and statistical approaches to language* (pp. 1–26). MIT Press.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton & Co.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied NLP* (pp. 136–148). Austin, TX.

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, *14*(1), 31–39.

Klavans, J. L., & Resnik, P. (Eds.). (1996). *The balancing act: Combining symbolic and statistical approaches to language.* MIT Press.