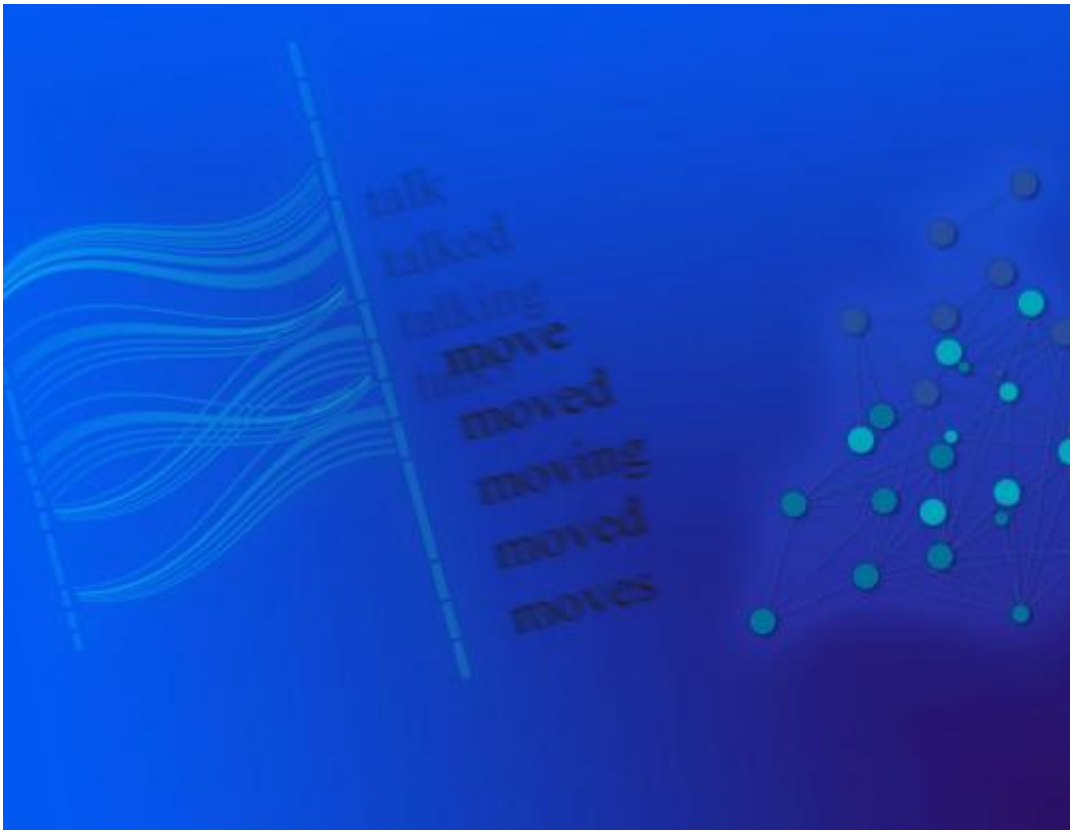
Linguists tackle computational analysis of grammar

February 26 2015, by Benjamin Recchie



The University of Chicago's Research Computing Center is helping linguists visualize the grammar of a given word in bodies of language containing millions or billions of words. Credit: Ricardo Aguilera/Research Computing Center

Children don't have to be told that "cat" and "cats" are variants of the same word—they pick it up just by listening. To a computer, though,

they're as different as, well, cats and dogs. Yet it's computers that are assumed to be superior in detecting patterns and rules, not 4-year olds. John Goldsmith, the Edward Carson Waller Distinguished Service Professor of Linguistics and Computer Science, and graduate student Jackson Lee are trying to, if not to solve that puzzle definitively, at least provide the tools to do so.

Studying natural language morphology has both practical and theoretical aspects. Theoretically, linguists and cognitive scientists have long sought a better understanding of how humans learn language. "Computational modeling of how natural language morphology may be learned from raw text is an explicit attempt to answer this question," said Lee. And practically, better understanding of natural language morphology can lead to better designed human-machine interfaces and a better way to search large databases.

"We are trying to do computationally what linguists have always done," explained Goldsmith. "Collect large amounts of texts in a language, and produce grammatical analyses of the language. We would like to understand that process of what we"—humans and human linguists—"do so well that we can implement it computationally."

To provide examples for their analysis, Goldsmith and Lee used standard bodies of written language called corpora. Each corpus contains millions, sometimes billions, of words, taken from many different genres of writing. (The Brown corpus, the first of its kind in American English, contained roughly one million words; the Google N-gram corpus contains 155 billion words.) Their combined data set was far too big to be handled on a desktop computer. Instead, they turned to the Research Computing Center and the Midway supercomputing cluster for help. RCC consultants also helped them to make better use of Midway's multiple cores by helping them to parallelize their algorithms.

RCC consultants also helped Lee and Goldsmith to visualize their results. "A typical scenario for us is that, given some raw data, we have some intuition about certain patterns in the data, and we collaborate with RCC to create visualization tools to display data in a way that enables us to explore these patterns." Lee said. He gave the example of the query word "going": The visualization showed what words occur most frequently on the left and right of it in a natural language corpus.

"The construction of this visualization tool grew out of the observation that overall word distribution patterns are sensitive to the specific distribution of individual words, and we need a tool to 'see' what the grammar of a given word really looks like," Lee added. Lee and Goldsmith demonstrated this work in a poster presented at this past year's Mind Bytes symposium, where it won a special award from the judges for novel uses of computational resources.

Lee and Goldsmith are taking their work and developing it into an integrated research and visualization tool. "This includes not only the suite of the visualization tools developed, but also implementations of algorithms and ideas—both from us and other researchers—with regard to the unsupervised learning of linguistic structure," said Lee. The final product will allow different research groups to visualize their results and compare their methods.

But beyond just the computational problem, Goldsmith sees a deeper question waiting to be answered. Philosophers and linguists have long argued about whether a language can only be learned by understanding the meaning of the sentences that make it up. "At the end of the day," said Goldsmith, "language exists with the function of organizing and communicating meaning. But is it possible to define and detect grammatical structure even before knowing the meaning in a text?"

Provided by University of Chicago