CHICAGOLAWBULLETIN.COM
TUESDAY, SEPTEMBER 5, 2017

Chicago Daily Law Bulletin®

Volume 163, No. 163
Serving Chicago's legal community for 162 years

Using big data to decode lexical enigmas

If you are old enough to remember "The Summer of Love," you also remember going through a law library Shepardizing cases.

In addition to the bound volumes of Shepard's, there were at least 27 different pocket parts, seven of which — by law — had to be missing. And one of the missing pocket parts — again by law — had to show that your case had been recently overruled.

It is hard to explain to younger lawyers how wonderful it is now to Shepardize with a keystroke on Lexis or Westlaw. Computers have obviously revolutionized legal research.

But computers are now transforming yet another area of law: statutory construction. If you have not heard of "corpus linguistics," it is time you did.

The best place to start is Thomas R. Lee and Stephen C. Mouritsen's article, "Judging Ordinary Meaning," which will be published in a forthcoming issue of the Yale Law Journal. (It's available for free download at bit.ly/2wF4G4u.)

The article's thesis is that the field of linguistics has established principles and methods for determining the meaning of words. When judges inquire as to the "ordinary meaning" of a word in a statute, they are posing a question that can be answered through empirical means.

And since linguists have developed computer-driven methods for answering such a question, it is time for judges and lawyers to do the same.

Lexicographers now rely on electronic "corpora" — large bodies or databases of naturally occurring language — in order to gather linguistic data. This allows them to "view a more complete range of potential uses of a given word appearing in a particular semantic environment."

That is, corpus linguistics determines the meaning of a word by looking at "real-world language in its natural habitat in books, magazines, newspapers, and even transcripts of spoken language."

Want an example? Let's take H.L.A. Hart's classic statutory interpretation problem concerning a sign: "No vehicles allowed in park." Hart believed the sign plainly forbade people from driving cars into the park. But he asked whether the vehicle prohibition would also include bicycles, roller skates and toy cars. In short, are items in those three categories "vehicles" under the meaning of the statute?

Lee and Mauritsen provide an example of how corpus linguistics would approach the question of whether a bicycle is a forbidden "vehicle." They use what is known as the N.O.W. (News on the Web) Corpus, which is a database of 3.7 billion words from newspapers and magazines during the past seven years.

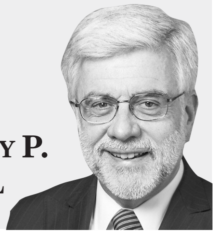
If you have not heard of "corpus linguistics," it is time you did.

They also utilize COHA, which is the Corpus of Historical American English. This contains more than 400 million words of text from the past two centuries.

The first step is the use of "lexical collocation." The collocation function of the database search is to determine "the words that are statistically most likely to appear in the same context as 'vehicle.'"

CRIMINAL PROCEDURE

TIMOTHY P. O'NEILL



Timothy P. O'Neill is the Edward T. and Noble W. Lee Chair in Constitutional Law for 2014-15 at The John Marshall Law School in Chicago. Readers are invited to visit his Web log and archives at jmls.edu/oneill.

Of the 50 most common collocates, most of them are related to automobiles, e.g., motor, car, traffic, fuel, and driving. On the other hand, "bicycle" does not appear among the collocates of "vehicle."

The second step is the use of "concordance data." This takes finds entire phrases in which the word "vehicle" is used in print. Lee and Mauritsen's review of 100 randomly chosen phrases containing "vehicle" showed that 91 were references to automobiles. They did not find a single reference to a bicycle.

The third step is to look not only at "vehicle" in general, but rather "vehicle" when it is used within the context of a "park." Here the majority of the references were to automobiles. Again, they could find no mention of bicycles.

The final step is the use of what is known as the KWIC function: "Key Words in Context." This looks for phrases which contain both the words "vehicle" and "bicycle" in order to determine their relationship.

Use of KWIC determined that bicycle was often used in contrast to vehicle, e.g., "There were 68 collisions between bicycles, pedestrians, and vehicles."

The conclusion? Based on empirical evidence, although "bicycle" is certainly a possible sense of a "vehicle," it is not a common meaning. It therefore seems doubtful that the prohibition "No vehicles allowed in park" would include bicycles.

Lee and Mauritsen admit that corpus linguistics is not a cure-all for all legal interpretive problems. It specifically deals with solving problems of "lexical ambiguity," i.e., where there is a question of two competing meanings of a term in the statutory text.

It has no application to "semantic" or "structural ambiguity;" regarding the phrase "Mary saw the man with the telescope," corpus linguistics cannot determine whether this means Mary used the telescope or the man was holding the telescope.

Moreover, they concede that judges, lawyers, and linguists will need to agree on a set of "best practices" in using this in law.

There needs to be consensus on what is the best corpus to use for a specific kind of ambiguity; standards on the appropriate sample size for a specific search; standards regarding appropriate search terms and search methods; and identification of suitable coding methods.

Lee and Mauritsen admit that further work needs to be done. But they do insist that judges and lawyers are already linguists, whether they know it or not. And if you are going to engage in linguistic analysis, you need to begin thinking and analyzing like a linguist. Corpus linguistics is a start.