Why context matters in the long and short of words: Researchers improve 75-year-old language theory

June 20 2011, By Bobbie Mixon

In 1935, Harvard University linguist George Kingsley Zipf asserted that "the magnitude of words tends, on the whole, to stand in an inverse, not necessarily proportionate, relationship to the number of occurrences." In other words, short words are used more frequently than long ones. Now, cognitive scientists at the Massachusetts Institute of Technology demonstrated a substantial improvement to Zipf's law. Credit: Adrian Apodaca, National Science Foundation

(Medical Xpress) -- Do you ever wonder about the stuff that makes up words? Why is a word a word, what goes into forming it, what's its history or why is it long or short? Scientists at the Massachusetts Institute of Technology do.

Steven Piantadosi, Harry Tily and Edward Gibson study words for MIT's Department of Brain and Cognitive Sciences to understand how humans

think and communicate.

Recently, they put a well-established, 75-year-old language theory to the test and found it had room for improvement. At issue was something called Zipf's law, an empirical scientific principle that says word length is primarily determined by frequency of use.

In 1935, Harvard University linguist George Kingsley Zipf asserted "the magnitude of words tends, on the whole, to stand in an inverse, not necessarily proportionate, relationship to the number of occurrences." In other words, short words are used more than long ones.

"One widely known and apparently universal property of human language is that frequent words tend to be short," the researchers write in their report. They note short words are used to make communication more efficient than what can be had with frequent use of longer words.

This is because of pressure for communicative efficiency, Zipf surmised. It would be impractical to ask everyone at a Thanksgiving dinner whether they would like a bowl of soup using a 15-letter word for "of," for example.

In the Brown University Standard Corpus of Present-Day American English, which contains about two million words of text, "of" is the fourth most commonly used word. Meanwhile, "the" is used more in writing than any other word in the English language. In fact, a list of the top 100 most frequently used words contains words such as "be," "on," "have," "with," "who," and "some," all very short words.

But the cognitive scientists at MIT demonstrated a substantial improvement to Zipf's law. They showed that across 10 languages the predictability of what a person says is a more important determinant of word length than how often he or she says it.

Word length actually comes down to the amount of information it contains

The goal of the research was to compare Zipf's word frequency theory to Piantadosi and colleagues' word predictability theory--the idea that the average amount of information a word conveys in context--its predictability--determines word length.

Using an Internet database, the researchers studied how often all possible sequences of two, three or four word combinations occur together in order to estimate how predictable any word is when it's typically written.

By knowing this, they could determine whether context and predictability were better determinants of word length than frequency of use.

"For instance, in a context like ‘Monday night ____' the word ‘football' is very predictable and therefore conveys very little information," said Piantadosi, a cognitive scientist in the Ph.D. program at MIT and lead author of the study. "But, in a context like ‘I ate ____,' the missing word is very unpredictable, but conveys a lot of information."

The hypothesis was that average information contained in two, three or four word sequences should in part determine the length of words, either in letters or syllables, since that's how an optimal code would behave. In this example, "football" and the two words preceding it demonstrated the effect.

"The only way these effects can get in to the lexicon is if our linguistic systems, and the mechanisms of language change, are sensitive to communicative pressures," said Piantadosi.

The sequences of words that people use are coded--their letters, syllables, sounds, etc.--for efficient communication and are better predictors of word length than frequency alone, he said.

"This means word sequences provide efficient codes for the meanings they convey, relative to the statistical regularities in language," he said. "That's our claim."

Context matters for love, amour, liebe, amor and kärlek,

Love, amour, liebe, amor and kärlek all mean the same thing across different languages and all are about the same length, which according to Zipf is what should be expected if they were similarly predictable or informative. But the MIT researchers stress it's the words before and after a particular word that determines how often the particular word is used, not length.

True, the word for strong fondness is very short, but how frequently do people say it, what are the circumstances when they do and how predictable is the information conveyed when it's said? Saying "I love you" is quite different from saying "I love chicken." For a word like "love," context matters.

The research results held across all but one of the languages studied: Czech, Dutch, English, French, German, Italian, Portuguese, Romanian, Spanish and Swedish, with German being the outlier.

"I was surprised that we found effects in so many languages," said Piantadosi. "I would have thought that differences in morphology, or word structure, might have swamped our effects in many languages, but this doesn't appear to be the case."

Why the most frequently used words are short

The research findings also provide an improved explanation as to why the most often used words are short--because they tend to be predictable, meaning many short words, on average, convey relatively little information. Of the top 100 words, many are "function words," whose main purpose is to join words together such as--"with," "from" and "over." By themselves, these words give the reader or listener a very small amount of data.

The researchers also found short words must be paired with other familiar words to derive context and convey information. This is because many times words occurring after well-known sequences of other words are the most predictable and contain the least information; for example "a ton of fun," is a well known sequence of words that conveys very little information. But words that have a little association to the words preceding them contain more information; for example, "a ton of butter."

A final word

The research revealed that people communicate through at least an approximately optimal code for meaning, said Piantadosi. "Lexicons are not arbitrary in the sense of being completely random. Instead, they are well-structured for communication, given the patterns of word sequences people typically use."

The problem with the traditional method of only looking at word frequency is that it merely involves counting words in isolation and does not consider the regular dependencies between words.

The research is published in the *Proceeding of the National Academy of Sciences* in an article titled "Word lengths are optimized for efficient communication." The National Science Foundation's Division of

Behavioral and Cognitive Sciences funds the research.

More information: Word lengths are optimized for efficient communication: www.pnas.org/content/early/201 … /1012551108.abstract

Provided by National Science Foundation