What determines the length of words? MIT researchers say they know

February 10 2011

Graphic: Christine Daniloff

Why are some words short and others long? For decades, a prominent theory has held that words used frequently are short in order to make language efficient: It would not be economical if "the" were as long as "phenomenology," in this view. But now a team of MIT cognitive scientists has developed an alternative notion, on the basis of new research: A word's length reflects the amount of information it contains.

"It may seem surprising, but word lengths are better predicted by information content than by frequency," says Steven Piantadosi, a PhD

candidate in MIT's Department of Brain and Cognitive Sciences (BCS), and the lead author of a paper on the subject that evaluates word use in 11 languages. The paper was published online last month in the *Proceedings of the National Academy of Sciences (PNAS)*.

The notion that frequency of use engenders shorter words stems from work published by Harvard scholar George Zipf in the 1930s. The Zipf idea, Piantadosi notes, has an intuitive appeal to it, but only offers a limited explanation of word lengths. "It makes sense that if you say something over and over again, then you want it to be short," Piantadosi says. "But there is a more refined communications story to be told than that. Frequency doesn't take into account dependencies between words."

That is, many words typically appear in predictable sequences along with other words. Short words are not necessarily highly frequent; more often, the researchers found, short words do not contain much information by themselves, but appear with strings of other familiar words that, as an ensemble, convey information.

In turn, this clustering of short words helps "smooth out" the flow of information in language by forming strings of similar-sized language packets, which creates an efficiency of its own — albeit not exactly the one Zipf envisioned. "If you take the view that people should be trying to communicate efficiently, you get this uniform rate," adds Piantadosi; whether delivered through clusters of shorter words or through individual longer words carrying greater information, language tends to convey information at consistent rates.

Written in the script

Piantadosi conducted the study along with Edward Gibson, a professor in BCS who also has a joint appointment in the Department of Linguistics, and Harry Tily, a postdoctoral associate in BCS. In the paper, the MIT

researchers studied an enormous data set of online documents posted by Google. Since the documents included a lot of Internet-specific character sequences not comprising words — think "www" — the team began its search by cataloguing texts from Open Subtitles, a database of movie translations, and searched for the words used in those documents when mining the larger Google database. “Movie subtitles are words used naturalistically, so we took words used frequently in that data set and pulled their statistics from Google,” explains Piantadosi. The 11 languages in the study are all European.

To evaluate how much information is contained in a word, the researchers defined information as existing in an inverse relationship to the predictability of words. That is, the words most often occurring after familiar sequences of two, three or four other words — such as the “eat” in “you are what you eat” — contain the least information individually. By contrast, words whose appearances have a minimal relationship to the words preceding them — such as the “contagious” in “you are contagious” — contain, individually, more information. This principle is based on the highly influential work of former MIT information-theory pioneer Claude Shannon.

The MIT team found that 10 percent of the variation in word length is attributable to the amount of information contained in those words — not a high figure by itself, but one about three times as large as the variation in word length attributable to frequency, the notion Zipf championed. For English words, 9 percent of the variation in length is due to amount of information, and 1 percent stems from frequency. It turns out, for instance, that words as disparate in length as “mind” and “organization” appear with virtually the same frequency. However, as Gibson acknowledges, “the data itself is noisy,” and there are counter-examples that do not necessarily support their thesis; for instance, “menu” and “selection” have about the same informational content.

Colleagues believe the study's new insight about the mechanics of language will prove important over time. "This is exciting work," says Roger Levy, an assistant professor in the Department of Linguistics at the University of California, San Diego. In Levy's view, the paper answers an important objection to Zipf's law lodged by George Miller, a psychologist at Princeton University. As Miller pointed out, any random language generator using a space key — the proverbial monkeys on a typewriter — would also create language patterns in which shorter strings of characters appear most frequently.

By contrast, the current paper, while offering an alternative view of efficiency to the one Zipf held, does imply that word length has a non-random basis. "The notion of monkeys on a typewriter can't explain these findings," adds Levy.

Still, the researchers acknowledge there is much more work to be done in this area of language studies. Piantadosi, for one, is using similar data-mining techniques to study the role of ambiguity in language, studying how the meaning of words with multiple potential definitions becomes clarified by the presence of frequently appearing words around them. He hopes to publish results about the subject as a follow-up to the current *PNAS* paper.

Provided by Massachusetts Institute of Technology