Unzipping Zipf's Law: Solution to a century-old linguistic problem

10 August 2017

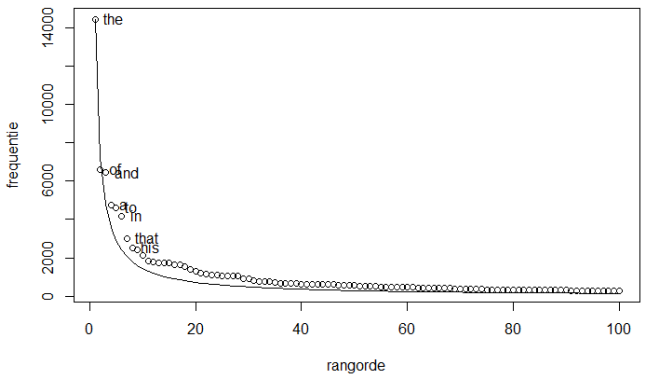

Figure 1. Zipfian distribution of the frequency (vertical axes) and the rank in the frequency table (horizontal axes) of the first hundred words of Melvilles Moby Dick. The line was predicted by Zipf's law, and the dots depict the actual word frequencies in the text. Credit: Radboud University

Did you know that in every language, the most frequent word occurs twice as often as the second most frequent word? This phenomenon called 'Zipf's law' is more than one century old, but until now, scientists have not been able to elucidate it exactly. Sander Lestrade, a linguist at Radboud University in The Netherlands, proposes a new solution to this notorious problem in *PLOS ONE*.

Zipf's law describes how the frequency of a word in natural language, is dependent on its rank in the frequency table. So the most frequent word occurs twice as often as the second most frequent word, three times as often as the subsequent word, and so on until the least frequent word (see Figure 1). The law is named after the American linguist George Kingsley Zipf, who was the first who tried to explain it around 1935.

Biggest mystery in computational linguistics

"I think it's safe to say that Zipf's law is the biggest mystery in computational linguistics," says Sander Lestrade, linguist at Radboud University in Nijmegen, the Netherlands. "In spite of decades of theorizing, its origins remain elusive." Lestrade now shows that Zipf's law can be explained by the interaction between the structure of sentences (syntax) and the meaning of words (semantics) in a text. Using computer simulations, he was able to show that neither syntax or semantics suffices to induce a Zipfian distribution on its own, but that syntax and semantics 'need' each other for that.

"In the English language, but also in Dutch, there are only three articles, and tens of thousands of nouns," Lestrade explains. "Since you use an article before almost every noun, articles occur way more often than nouns." But that is not enough to explain Zipf's law. "Within the nouns, you also find big differences. The word 'thing', for example, is much more common than 'submarine', and thus can be used more frequently. But in order to actually occur frequently, a word should not be too general either. If you multiply the differences in meaning within word classes, with the need for every word class, you find a magnificent Zipfian distribution. And this distribution only differs a little from the Zipfian ideal, just like natural language does, as you can see in Figure 1."

Not only are predictions based on Lestrades new model completely consistent with phenomena found in natural language, his theory also holds for almost every language in the world, not only for English or Dutch. Lestrade: "I am overjoyed with this finding, and I am convinced of my theory. Still, its confirmation must come from other linguists."

More information: Sander Lestrade et al. Unzipping Zipf's law, *PLOS ONE* (2017). DOI: 10.1371/journal.pone.0181987

Provided by Radboud University