Words, more words ... and statistics

May 17 2016

Picking out single words in a flow of speech is no easy task and, according to linguists, to succeed in doing it the brain might use statistical methods. A group of SISSA scientists has applied a statistics-based method for word segmentation and measured its efficacy on natural language, in nine different languages, to discover that linguistic rhythm plays an important role. The study has just been published in the *Journal of Developmental Science*.

Have you ever racked your brains trying to make out even a single word of an uninterrupted flow of speech in a language you hardly know at all? It is naïve to think that in speech there is even the smallest of pauses between one word and the next (like the space we conventionally insert between words in writing): in actual fact, speech is almost always a continuous stream of sound. However, when we listen to our native language, word "segmentation" is an effortless process. What are, linguists wonder, the automatic cognitive mechanisms underlying this skill? Clearly, knowledge of the vocabulary helps: memory of the sound of the single words helps us to pick them out. However, many linguists argue, there are also automatic, subconscious "low-level" mechanisms that help us even when we do not recognise the words or when, as in the case of very young children, our knowledge of the language is still only rudimentary. These mechanisms, they think, rely on the statistical analysis of the frequency (estimated based on past experience) of the syllables in each language.

One indicator that could contribute to segmentation processes is "transitional probability" (TP), which provides an estimate of the

likelihood of two syllables co-occurring in the same word, based on the frequency with which they are found associated in a given language. In practice, if every time I hear the syllable "TA" it is invariably followed by the syllable "DA", then the transitional probability for "DA", given "TA", is 1 (the highest). If, on the other hand, whenever I hear the syllable "BU" it is followed half of the time by the syllable "DI" and half of the time by "FI", then the transitional probability of "DI" (and "FI"), given "BU", is 0.5, and so forth. The cognitive system could be implicitly computing this value by relying on linguistic memory, from which it would derive the frequencies.

The study conducted by Amanda Saksida, research scientist at the International School for Advanced Studies (SISSA) in Trieste, with the collaboration of Alan Langus, SISSA research fellow, under the supervision of SISSA professor Marina Nespor, used TP to segment natural language, by using two different approaches.

Based on rhythm

Saksida's study is based on the work with corpora, that is, bodies of texts specifically collected for linguistic analysis. In the case at hand, the corpora consisted of transcriptions of the "linguistic sound environment" that infants are exposed to. "We wanted to have an example of the type of linguistic environment in which a child's language develops", explained Saksida, "We wondered whether a low-level mechanism such as transitional probability worked with real-life language cues, which are very different from the artificial cues normally used in the laboratory, which are more schematic and free of sources of 'noise'. Furthermore, the question was whether the same low-level cue is equally efficient in different languages". Saksida and colleagues used corpora of no less than 9 different languages, and to each they applied two different TP-based models.

First they calculated the TP values for each point of the language flow for all of the corpora, and then they "segmented" the flow using two different methods. The first was based on absolute thresholding: a certain fixed reference TP value was established below which a boundary was identified. The second method was based on relative thresholding: the boundaries corresponded to the locally lowest TP function.

In all cases, Saksida and colleagues found that transitional probability was an effective tool for segmentation (49% to 86% of words identified correctly) irrespective of the segmentation algorithm used, which confirms TP efficacy. Of note, while both models proved to be quite efficient, when one model was particularly successful with one language, the alternative model always performed significantly worse.

"This cross-linguistic difference suggests that each model is better suited than the other for certain languages and viceversa. We therefore conducted further analyses to understand what linguistic features correlated with the better performance of one model over the other", explains Saksida. The crucial dimension proved to be linguistic rhythm. "We can divide European languages into two large groups based on rhythm: stress-timed and syllable-timed". Stress-timed languages have fewer vowels and shorter words, and include English, Slovenian and German. Syllable-timed languages contain more vowels and longer words on average, and include Italian, Spanish and Finnish. The third rhythmic group of languages does not exist in Europe and is based on "morae" (a part of the syllable), such as Japanese. This group is known as "mora-timed" and contains even more vowels than syllable-timed languages.

The absolute threshold model proved to work best on stress-timed languages, whereas relative thresholding was better for the mora-timed ones. "It's therefore possible that the cognitive system learns to use the

segmentation algorithm that is best suited to one's native language, and that this leads to difficulties segmenting languages belonging to another rhythmic category. Experimental studies will clearly be necessary to test this hypothesis. We know from the scientific literature that immediately after birth infants already use rhythmic information, and we think that the strategies used to choose the most appropriate segmentation could be one of the areas in which information about rhythm is most useful".

The study is in fact unable to say whether the cognitive system (of both adults and children) really uses this type of strategy. "Our study clearly confirms that this strategy works across a wide range of languages", concludes Saksida. "It will now serve as a guide for laboratory experiments."

More information: Amanda Saksida et al, Co-occurrence statistics as a language-dependent cue for speech segmentation, *Developmental Science* (2016). DOI: 10.1111/desc.12390

Provided by International School of Advanced Studies (SISSA)

Citation: Words, more words ... and statistics (2016, May 17) retrieved 19 April 2024 from https://phys.org/news/2016-05-words-statistics.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.