TechXplore

Computer system analyzes differences in the syntax of languages

November 10 2022, by Myrthe Timmers

Credit: Pixabay/CC0 Public Domain

For decades, linguists have racked their brains over the question of precisely how the syntax of various languages is different. Ph.D. candidate Martin Kroon has developed a computer system that brings us closer to finding an answer. His Ph.D. defense is on 10 November.

Knowing about the similarities and differences between languages will bring us one step closer to understanding how our brains work. After all, discovering a structure that is shared by different languages could tell us a great deal about how the brain handles language. Until now, however, it has proved difficult to identify all the ways in which languages are the same or different.

"This is all done manually, but there are an awful lot of languages and basically an infinite number of sentences you can generate in them," Kroon explains. This means that there's a risk of bias. "You have to select in advance what you're going to compare, which can cause you to overlook things or conversely to confirm things that don't occur very often at all."

Compressing language

Kroon therefore decided to take a different approach. A computer system should make it possible to compare different languages on a larger scale. "I mainly used transcripts of EU meetings, because they're translated into all the European Union languages," he says, and then explains how he applied two methods to the data.

"First, I was impressed by the Minimum Description Length (MDL) principle. This is actually a matter of compression, the same as you do on your computer: how do you make big data as small as possible, so that they fit into a zip file? To do this, MDL searches for patterns that occur frequently but are not too long. In Dutch, for example, this could be 'article+noun.' This pattern is easy to compress and you won't find it in Czech, for example, because Czech doesn't have articles."

He found that the system worked. Patterns in the transcripts emerged, indicating syntactic similarities and differences. At the same time, however, the computer would often find differences that on closer inspection had very little to do with syntax.

"Some texts were translated manually, so you couldn't compare them syntactically any more," says Kroon. "For instance, the original English 'to the matter at hand' was translated into Dutch as 'en nu het eigenlijke onderwerp' (= 'and now the actual subject'). This means the same thing, but it's completely different in terms of syntax and structure."

Projecting English onto Hungarian

Moreover, the way in which the languages were described linguistically was not always helpful: descriptions of linguistic phenomena used in Dutch could not be found in Czech and vice versa. And, for instance, the Dutch "te" as in "te doen" (= "to do") was structurally tagged as a preposition, while its English counterpart "to" was structurally tagged as a particle. Or more arbitrarily, the European Union was often tagged in Czech as "adjective+noun," while in English it was labeled "proper noun."

"In the second test, I therefore projected the annotations of one language onto another, non-annotated language," says Kroon. "I knew too much about Czech by then, so I used Hungarian for the second test. First, we had to work out which words are each other's counterpart in sentences, which then allowed us to say: this is the finite verb in English, then this is probably the finite verb in Hungarian too."

Meanwhile, a Hungarian syntax specialist manually compiled a list of differences between English and Hungarian. Ideally, the software would find the same characteristic similarities and differences. "That didn't quite work out," Kroon has to admit. "We found confirmation for many of the hypotheses that I'd formulated on the basis of the software. But at the same time we weren't able to find all the characteristic differences. So my results can mostly give linguists a push in the right direction: try having a look here, because these might be interesting patterns. But completely automatic? As yet, we still need human interpretation too much for that."

More research is therefore needed. And Kroon sees this as definitely worthwhile. "All research starts with a question, and that question can only exist because we can put our thoughts into words. In my view, this means that researching language is just as important as everything else."

Provided by Leiden University

Citation: Computer system analyzes differences in the syntax of languages (2022, November 10) retrieved 16 April 2024 from
https://techxplore.com/news/2022-11-differences-syntax-languages.html