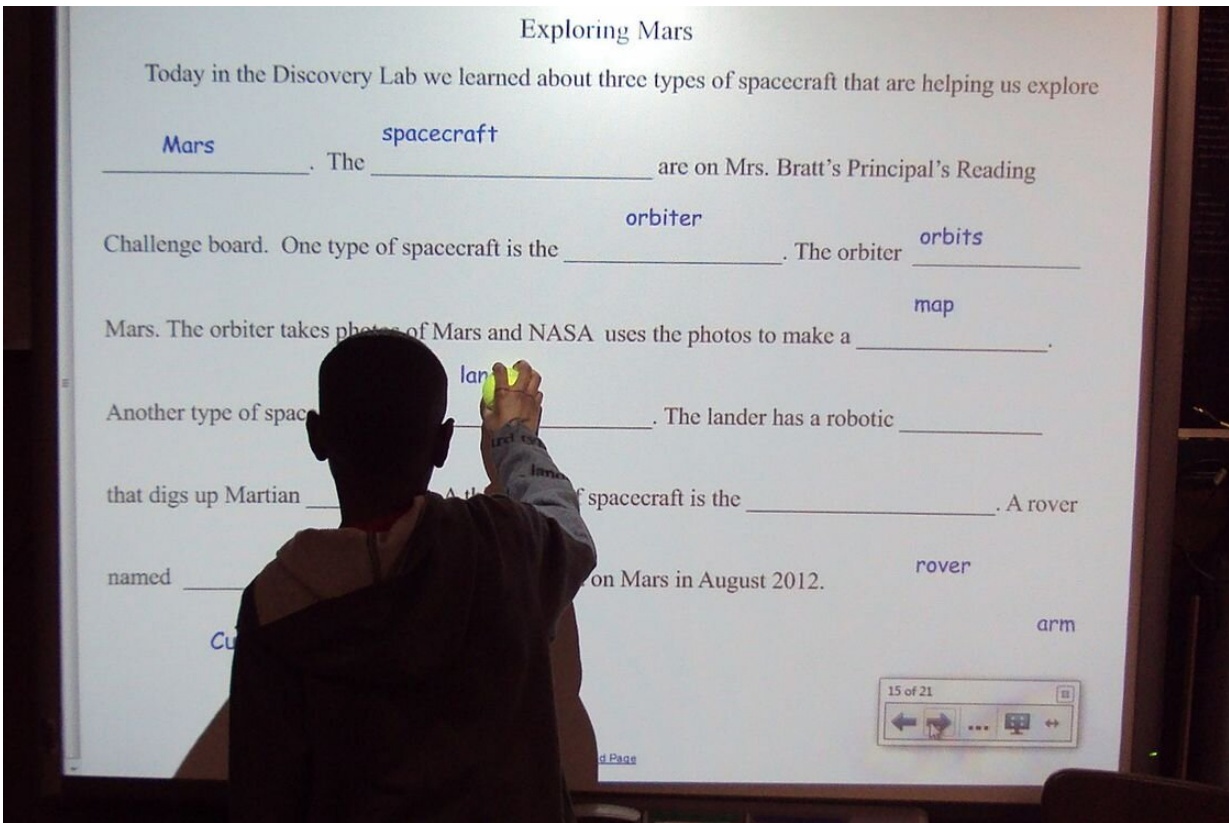# Predicting words' grammatical properties helps us read faster

February 16 2021

Student doing a cloze task on a smartboard. Credit: Wikimedia Commons

Psycholinguists from the HSE Centre for Language and Brain found that when reading, people are not only able to predict specific words, but also words' grammatical properties, which helps them to read faster.

Researchers have also discovered that predictability of words and grammatical features can be successfully modeled with the use of neural networks. The study was published in the journal *PLOS ONE*.

The ability to predict the next word in another person's speech or in reading has been described by many psycho- and neurolinguistic studies over the last 40 years. It is assumed that this ability allows us to process the information faster. Some recent publications on the English language have demonstrated evidence that while reading, people can not only predict specific words, but also their properties (e.g., the part of speech or the semantic group). Such partial prediction also helps us to read faster.

In order to access predictability of a certain word in a context, researchers usually use cloze tasks, such as "The cause of the accident was a mobile phone, which distracted the _____." In this phrase, different nouns are possible, but driver is the most probable, which is also the real ending of the sentence. The probability of the word "driver" in the context is calculated as the number of people who correctly guessed this word over the total number of people who completed the task.

The other approach for predicting word probability in context is the use of language models that offer word probabilities relying on a big corpus of texts. However, there are virtually no studies that would compare the probabilities received from the cloze task to those from the language model. Additionally, no one has tried to model the understudied grammatical predictability of words. The authors of the paper decided to learn whether native Russian speakers would predict grammatical properties of words and whether the language model probabilities could become a reliable substitution to probabilities from cloze tasks.

The researchers analyzed responses of 605 native Russian speakers in

the cloze task in 144 sentences and found out that people can precisely predict the specific word in about 18% of cases. Precision of prediction of parts of speech and morphological features of words (gender, number and case of nouns; tense, number, person and gender of verbs) varied from 63% to 78%. They discovered that the neural network model, which was trained on the Russian National Corpus, predicts specific words and grammatical properties with precision that is comparable to people's answers in the experiment. An important observation was that the neural network predicts low-probability words better than humans and predicts high-probability words worse than humans.

The second step in the study was to determine how experimental and corpus-based probabilities impact reading speed. To look into this, the researchers analyzed data on eye movement in 96 people who were reading the same 144 sentences. The results showed that first, the higher the probability of guessing the part of speech, gender and number of nouns, as well as the tense of verbs, the faster the person read words with these features.

The researchers say that this proves that for languages with rich morphology, such as Russian, prediction is largely related to guessing words' grammatical properties.

Second, probabilities of grammatical features obtained from the neural network model explained reading speed as correctly as experimental probabilities. 'This means that for further studies, we will be able to use corpus-based probabilities from the language model without conducting new cloze task-based experiments,' commented Anastasiya Lopukhina www.hse.ru/en/staff/lopukhina, author of the paper and Research Fellow at the HSE Centre for Language and Brain.

Third, the probabilities of specific words received from the language model explained reading speed in a different way as compared to

experiment-based probabilities. The authors assume that such a result may be related to different sources for corpus-based and experimental probabilities: corpus-based methods are better for low-probability words, and experimental ones are better for high-probability ones.

'Two things have been important for us in this work. First, we found out that reading native speakers of languages with rich morphology actively involve grammatical predicting,' Anastasiya Lopukhina said. 'Second, our colleagues, linguists and psychologists who study prediction got an opportunity to assess word probability with the use of language model: lm.ll-cl.org/. This will allow them to simplify the research process considerably'.

  More information: Anastasiya Lopukhina et al. Morphosyntactic but not lexical corpus-based probabilities can substitute for cloze probabilities in reading experiments, *PLOS ONE* (2021). DOI: 10.1371/journal.pone.0246133

Provided by National Research University Higher School of Economics