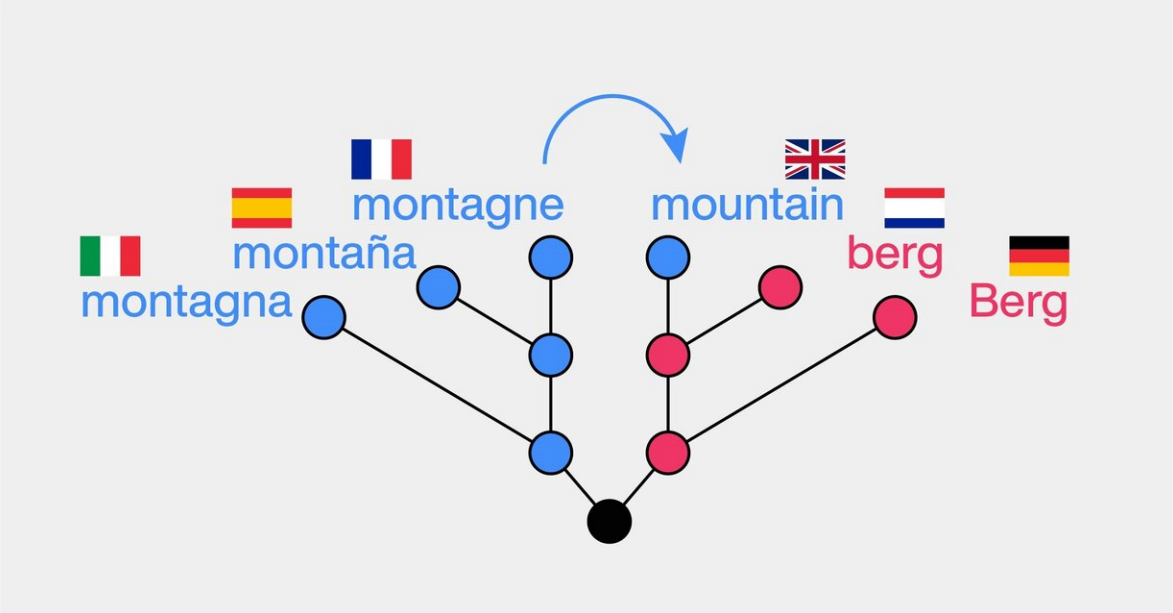
New study tests machine learning on detection of borrowed words in world languages

December 9 2020



Lexical borrowing is very widespread and may affect even those words that play an important role in our daily life. English 'mountain', for example, was borrowed from Old French, along with many other words. Credit: Johann-Mattis List, Hans Sell

Researchers from the Pontificia Universidad Católica del Perú and the Max Planck Institute for the Science of Human History have

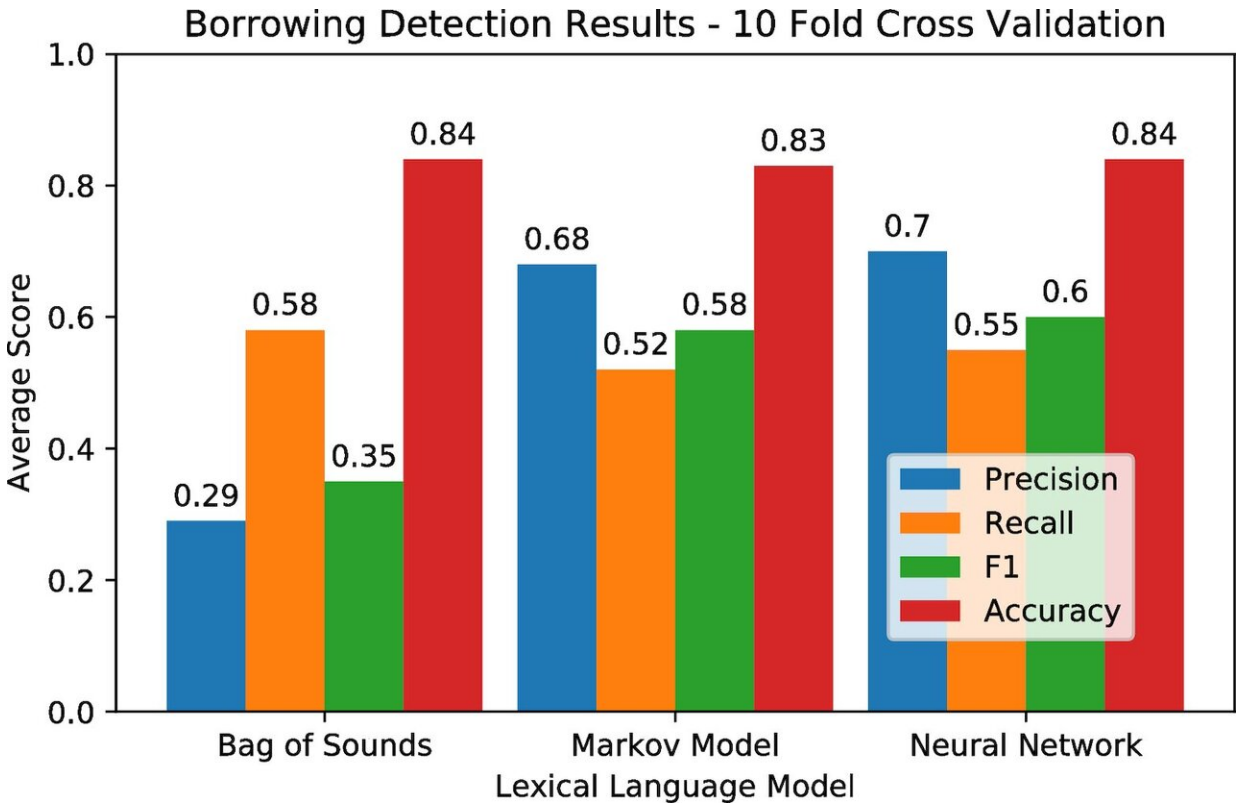
investigated the ability of machine learning algorithms to identify lexical borrowings using word lists from a single language. Results published in the journal *PLOS ONE* show that current machine-learning methods alone are insufficient for borrowing detection, confirming that additional data and expert knowledge are needed to tackle one of historical linguistics' most pressing challenges.

Lexical borrowing, or the direct transfer of words from one language to another, has interested scholars for millennia, as evidenced in Plato's Kratylos dialog, in which Socrates discusses the challenge imposed by borrowed words on etymological studies. In historical linguistics, lexical borrowings help researchers trace the evolution of modern languages and indicate cultural contact between distinct linguistic groups—whether recent or ancient. However, the techniques for identifying borrowed words have resisted formalization, demanding that researchers rely on a variety of proxy information and the comparison of multiple languages.

"The automated detection of lexical borrowings is still one of the most difficult tasks we face in computational historical linguistics," says Johann-Mattis List, who led the study.

In the current study, researchers from PUCP and MPI-SHH employed different machine learning techniques to train language models that mimic the way in which linguists identify borrowings when considering only the evidence provided by a single language: if sounds or the ways in which sounds combine to form words are atypical when comparing them with other words in the same language, this often hints to recent borrowings. The models were then applied to a modified version of the World Loanword Database, a catalog of borrowing information for a sample of 40 languages from different language families all over the world, in order to see how accurately words within a given language would be classified as borrowed or not by the different techniques.

In many cases the results were unsatisfying, suggesting that loanword detection is too difficult for machine learning methods most commonly used. However, in specific situations, such as in lists with a high proportion of loanwords or in languages whose loanwords come primarily from a single donor language, the teams' lexical language models showed some promise.



Comparison of the lexical language models used in the study. Credit: Miller et al., 2020

"After these first experiments with monolingual lexical borrowings, we can proceed to stake out other aspects of the problem, moving into multilingual and cross-linguistic approaches," says John Miller of PUCP,

the study's co-lead author.

"Our computer-assisted approach, along with the dataset we are releasing, will shed a new light on the importance of computer-assisted methods for language comparison and historical linguistics," adds Tiago Tresoldi, the study's other co-lead author from MPI-SHH.

The study joins ongoing efforts to tackle one of the most challenging problems in historical linguistics, showing that loanword detection cannot rely on mono-lingual information alone. In the future, the authors hope to develop better-integrated approaches that take multi-lingual information into account.

More information: *PLOS ONE* (2020). DOI: 10.1371/journal.pone.0242709

Provided by Max Planck Society