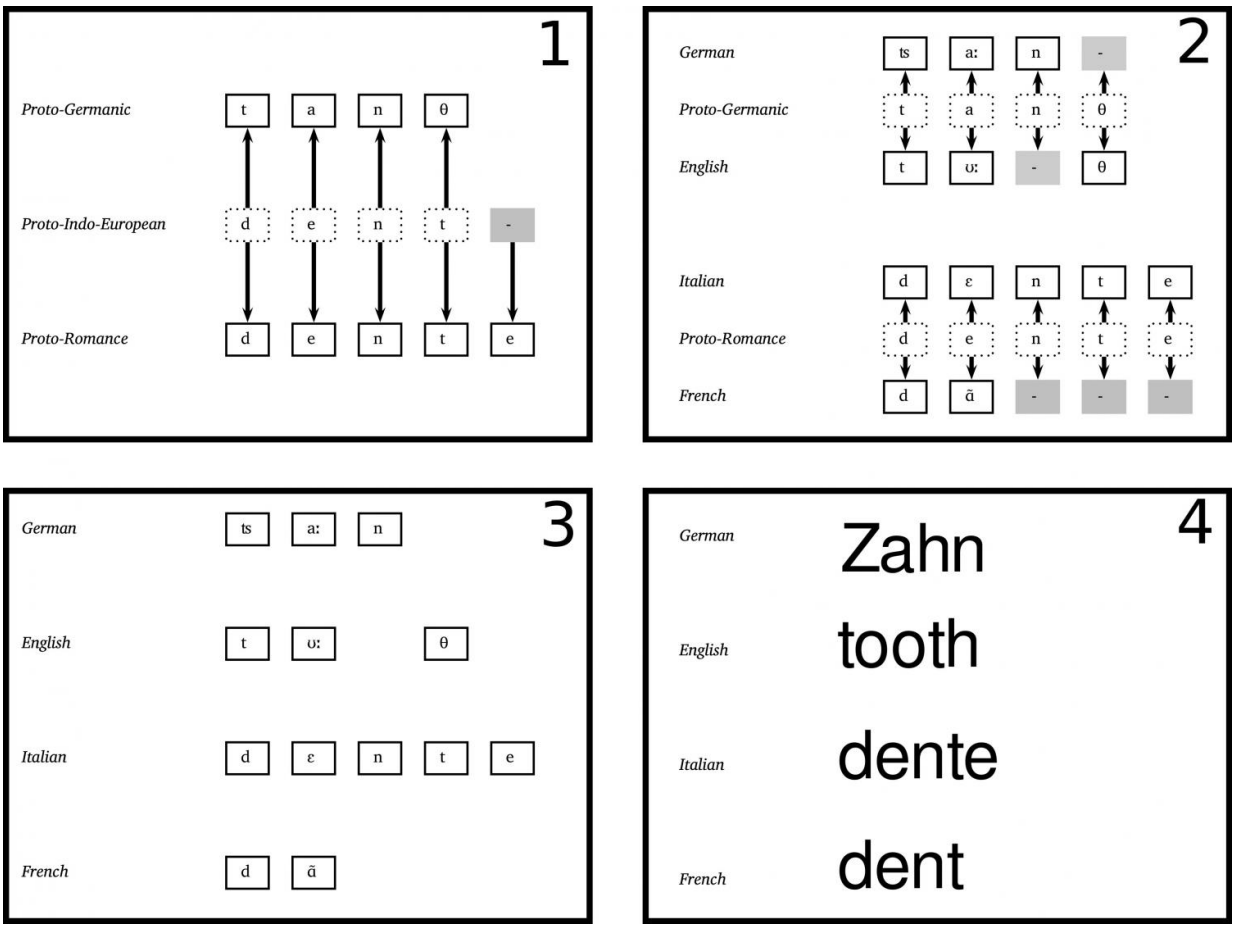
Computational methods applied to big datasets are compelling tools for historical linguistics

February 7 2017



Algorithms search huge datasets in order to determine the relationship among the languages spoken today. Even the comparison of individual words may give us hints about the past of our languages, as shown in the example, where the development of the word "tooth" in different Indo-European languages is

displayed. Credit: Johann-Mattis List

Digital approaches applied to big data play an increasingly important role in the humanities. However, there is skepticism about the accuracy and potential of computational methods for historical linguistics. A key task is the identification of etymologically related words (cognates) with a common ancestor, such as stone in English and Stein in German. Up to now, cognate detection is exclusively carried out by trained historical linguists who manually examine big datasets. This could change rather sooner than later, as a recent study by Johann-Mattis List, Simon Greenhill and Russell Gray from the Max Planck Institute for the Science of Human History has now revealed: The team has tested the capacity of different computational approaches to detect cognates – with striking success rates: The best-performing method could detect word relationships with an accuracy level of nearly 90%. This result not only confirms the potential of computational methodologies in the humanities, but also opens up exciting new pathways for future research in historical linguistics and human prehistory.

The comparison of different languages is a core task of historical linguistics. Language comparison allows linguists to trace the development of languages over thousands of years, long before writing systems or written records testified to the existence of languages. Words like tooth in English, Zahn in German, dente in Italian, and dent in French all go back to the same ancestor. Just as biologists reconstruct extinct species, and archaeologists reconstruct ancient societies, linguists can reconstruct the pronunciation of ancient from modern words and show which languages have developed from the same ancestor. Linguistic evidence therefore plays a crucial role in uncovering human prehistory.

While large digital collections of language data are becoming more abundant, only a tiny fraction of the more than 7000 languages spoken today has been thoroughly analyzed. This is not surprising, given that classical comparative studies in linguistics are still based on manual work by linguistic experts. "With the rapidly growing amounts of data, traditional methods are just reaching their practical limits", says Johann-Mattis List. Nevertheless, the need for historical language comparison is still vital: "In large parts of the world, like in New Guinea or South America, both the languages and the history of the human populations speaking them still remain crudely understudied", says List.

Big questions, growing data - and huge problems for computational analysis

Using computational approaches to analyze the large amounts of linguistic data in order to find answers to the big questions of human history and cultural evolution is appealing- and tricky. Unlike the careful linguistic analyses carried out by trained, experienced scholars, with detailed knowledge of specific languages, computer algorithms are blind to language-specific peculiarities and have to infer the parameters from the data that is fed to them. This shortcoming runs the risk of obtaining false results.

"Computational methods are often criticized for being a 'black-box'", says Simon Greenhill, second author of the study. "You may get a beautiful result, but you can't really evaluate its quality and reliability. What we really want to know is whether languages are related and which pieces of evidence actually support this inference".

In their study, the group directed by Russell Gray has tested the performance of different automated approaches varying in sophistication and complexity. The results were surprisingly good. "Our results were

quite accurate in most cases", says List. While some algorithms work really well under certain conditions, they may yield disappointing results under other circumstances. The best of the tested methods was a new approach which the team had developed specifically for their study. It detected cognates correctly and in agreement with expert judgments in 89.5% of all cases. "Contrary to the fear of many experts that automatic methods produce huge amounts of false positives we have actually found the inverse: If the algorithm says that two words are related, this is usually correct", Greenhill says.

The future is to combine algorithms and expert knowledge

Does this mean that machines will soon replace experts in the search for etymologically related words across the languages of the world? The Max Planck group does not suggest that this will be a successful strategy. Instead of exclusively computer-based approaches they favor computer-assisted strategies in which algorithmic methodologies are used to carry out preliminary analysis - the bulk of rough work - which can then be corrected by an expert. Russell Gray, director of the study, deems this to be only the beginning. "We have still not exhausted the full potential of computational methods in historical linguistics, and it is almost certain that future algorithms will bring us even closer to expert's judgments", he says. But computers will never be able to replace trained linguistic experts. Gray says: "Computational methods can take care of the repetitive and more schematic work. In this way, they will allow experts to concentrate on answering the interesting questions."

More information: Johann-Mattis List et al, The Potential of Automatic Word Comparison for Historical Linguistics, *PLOS ONE* (2017). DOI: 10.1371/journal.pone.0170046

Provided by Max Planck Society