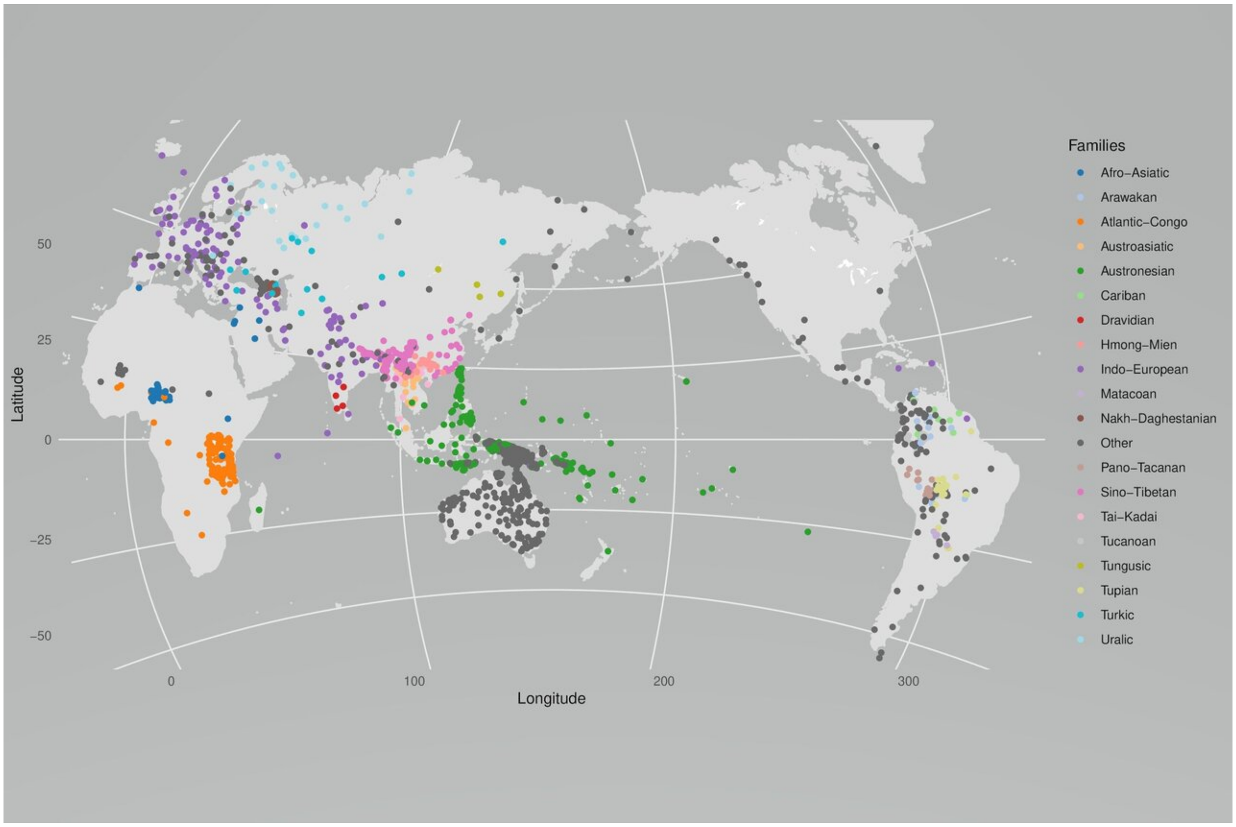
PHYS.ORG

CLICS: World's largest database of cross-linguistic lexical associations

January 13 2020



Global distribution of languages included in the CLICS3 release, identified by language family. Credit: S. J. Greenhill

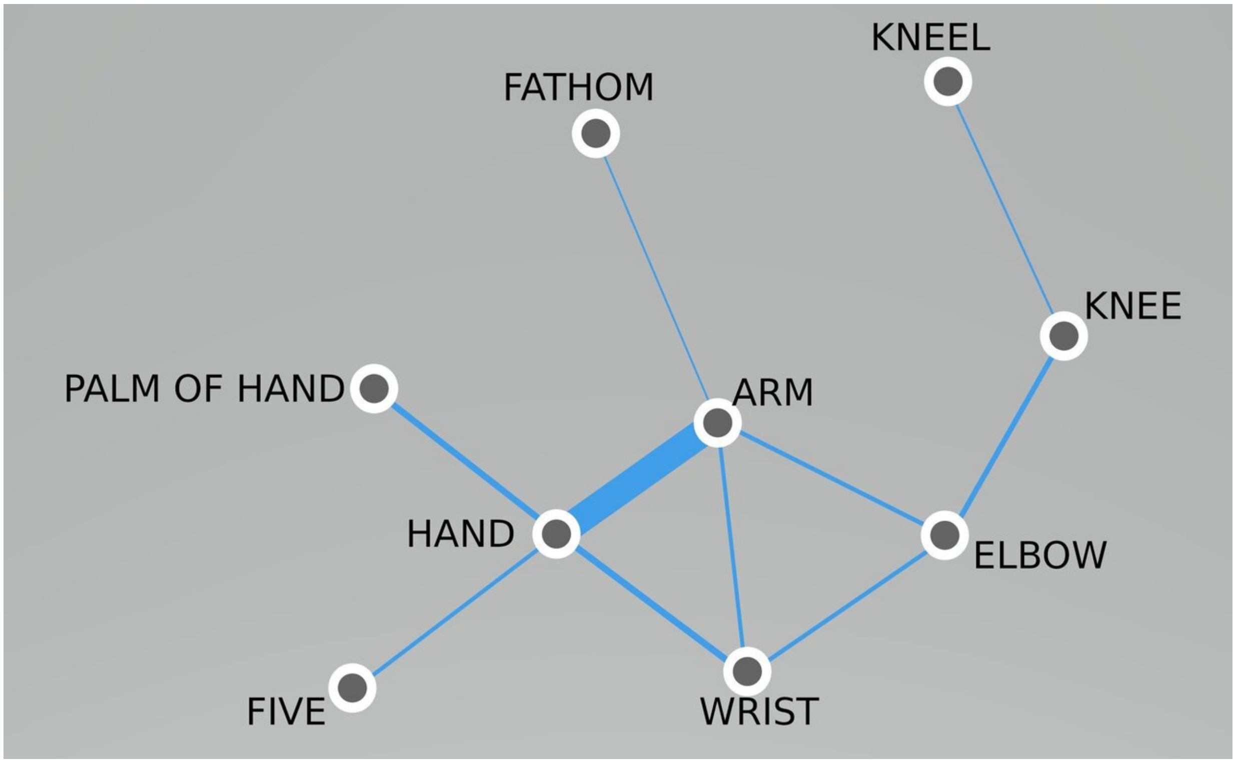
Every language has cases in which two or more concepts are expressed by the same word, such as the English word "fly," which refers to both the act of flying and to the insect. By comparing patterns in these cases, which linguists call colexifications, across languages, researchers can gain insights into a wide range of issues, including human perception, language evolution and language contact. The third installment of the CLICS database significantly increases the number of languages, concepts, and data sources available in earlier versions, allowing researchers to study colexifications on a global scale in unprecedented detail and depth.

With detailed computer-assisted workflows, CLICS facilitates the standardization of linguistic datasets and provides solutions to many of the persistent challenges in linguistic research. "While data aggregation was generally based on ad-hoc procedures in the past, our new workflows and guidelines for best practice are an important step to guarantee the reproducibility of linguistic research," says Tiago Tresoldi.

Effectiveness of CLICS demonstrated in research applications

The ability of CLICS to provide new evidence to address cutting-edge questions in psychology and cognition has already been illustrated in a recent study published in *Science*, which concentrated on the world-wide coding of emotional concepts. The study compared colexification networks of words for emotional concepts from a global sample of languages, and revealed that the meanings of emotions vary greatly across language families.

"In this study, CLICS was used to study differences in the lexical coding of emotion in languages around the world, but the potential of the database is not limited to emotion concepts. Many more interesting questions can be tackled in the future," says Johann-Mattis List.



Colexification network centered on the concepts "hand" and "arm." Credit: J.-M. List, T. Tresoldi

New standards and workflows allow for the reproducible harvesting of global lexical data

Building on the new guidelines for standardized data formats in cross-linguistic research, which were first presented in 2018, the CLICS team was able to increase the amount of data from 300 language varieties and 1200 concepts in the original database to 3156 language varieties and 2906 concepts in the current installation. The new version also guarantees the reproducibility of the data aggregation process, conforming to best practices in research data management. "Thanks to the new standards and workflows we developed, our data is not only

FAIR (findable, accessible, interoperable, and reproducible), but the process of lifting linguistic data from their original forms to our cross-linguistic standards is also much more efficient than in the past," says Robert Forkel.

The effectiveness of the workflow developed for CLICS has been tested and confirmed in various validation experiments involving a large range of scholars and students. Two different student tasks were conducted, resulting in the creation of new datasets and the progressive improvement of the existing data. Students were tasked with working through the different steps of data set creation described in the study, e.g. data extraction, data mapping (to reference catalogs), and identification of sources. "Having people from outside of the core team use and test your tools is essential and helps tremendously in fine-tuning all processes," says Christoph Rzymski.

With CLICS and its workflow being accessible to a wider audience, scholars cannot only directly contribute to the database in the future; they can also profit from the established machinery and start their own targeted collections. "The number of linguists who actively use our standards and workflows is constantly increasing. We hope that the release of this new version of CLICS will propagate them further," says Simon Greenhill.

More information: Christoph Rzymski et al. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies, *Scientific Data* (2020). DOI: 10.1038/s41597-019-0341-x

Provided by Max Planck Society

Citation: CLICS: World's largest database of cross-linguistic lexical associations (2020, January

13) retrieved 25 April 2024 from https://phys.org/news/2020-01-clics-world-largest-database-cross-linguistic.html