Max-Planck-Institut für Menschheitsgeschichte
Max Planck Institute for the Science of Human History

Strictly embargoed until 13 January, 2020 at 10:00 am GMT / 11:00 CET

CLICS: Världens största databas för tvärspråkliga lexikala associationer

Den senaste versionen av databasen sätter en ny standard för reproducerbar forskning genom ett pålitligt tillvägagångssätt för kvantitativa lingvistiska studier

Ett forskarteam under ledning av forskare vid Max Planck-institutet för mänsklighetens historia **har publicerat en ny version av Databasen för tvärspråkliga kolexifieringar (CLICS) som täcker lexikala associationer i mer än 3100 språk. Den nya databasen erbjuder tillgång till lexikala data av en aldrig tidigare skådad omfattning och förser användarna med ett detaljerat reproducerbart arbetsflöde för datainsamling, som gör det möjligt för forskare över hela världen att bidra till framtida versioner av databasen.**

Alla språk har ord som uttrycker flera begrepp, som exempelvis det engelska ordet *fly* vilket motsvarar två olika svenska ord – *flyga* (en handling) och *fluga* (en insekt). Sådana fall kallas kolexifieringar, och genom att jämföra mönstren i dessa mellan olika språk kan forskare få insikter om ett brett spektrum av frågor, bl. a sådana som berör människans perception, språkets utveckling och påverkan mellan språk. I den tredje versionen av databasen är antalet språk, begrepp och datakällor större än i tidigare versioner, vilket nu tillåter forskare att studera kolexifieringar i världens språk i större detalj och djup än tidigare.

De datorassisterade arbetsflödena i CLICS gör det lättare att standardisera dataset och erbjuder lösningar till flera av de återkommande utmaningarna i lingvistisk forskning. ”Medan den tidigare datainsamlingen mestadels baserades på ad-hoc-procedurer utgör våra nya arbetsflöden och riktlinjer ett viktigt steg mot att garantera att lingvistisk forskning ska kunna reproduceras” säger Tiago Tresoldi.

Effektiviteten i CLICS påvisad i forskning

CLICS förmåga att erbjuda ny evidens i sökande efter svar på banbrytande frågor inom psykologi och kognitionsvetenskap har redan visats i en studie om uttryck för emotionsbegrepp tvärs över världens språk, som nyligen publicerades i *Science*. Studien jämför kolexifieringsnätverk för ord som betecknar emotionsbegrepp i ett globalt urval av språk och demonstrerar att betydelsen hos emotionsorden kraftigt varierar mellan språkfamiljer.

”I denna studie har CLICS använts för att studera skillnaderna i lexikala uttryck för emotioner i språk över hela världen, men dess potential är inte begränsad till emotionsbegrepp. I framtiden kan flera andra intressanta forskningsfrågor hanteras med dess hjälp” säger Johann-Mattis List.

Nya standarder och arbetsflöden möjliggör reproducerbar insamling av lexikala data på global skala

Med utgångspunkt i de nya riktlinjerna för standardiserade dataformat i tvärspråklig forskning, som presenterades 2018 (DOI: 10.1038/sdata.2018.205) kunde CLICS-teamet utöka sina data från 300

språk och 1200 begrepp till 3156 språk och 2906 begrepp i den nuvarande versionen. Samtidigt garanterar den nya versionen att datainsamlingen är reproducerbar, i linje med bästa praxis inom forskningsdatahantering. ”Tack vare de nya standarder och arbetsflöden som vi utvecklat är inte bara våra data sökbara, tillgängliga, driftskompatibla och reproducerbara, utan själva processen att anpassa språkliga data från deras ursprungliga form till våra tvärspråkliga standarder är betydligt mer effektiv än tidigare” säger Robert Forkel.

Effektiviteten hos arbetsflödena inom CLICS har testats och bekräftats i flera olika valideringsexperiment som involverat ett brett spektrum av forskare och studenter. Två olika studentuppdrag har genomförts och lett till att nya dataset skapats och tidigare existerande data gradvis förbättrats. Studenterna fick arbeta med uppgifter relaterade till de olika steg i uppbyggnaden av data som beskrivs i studien, t ex datautvinning, datamappning (till referenskataloger), och identifiering av datakällor. ”Att ens redskap testas av människor utanför kärnan i forskningsgruppen är väsentligt och hjälper kollosallt vid finjusteringen av alla processer” säger Christoph Rzymski.

I och med att CLICS och dess arbetsflöden är tillgängliga för en bredare publik kan forskare bidra till databasen i framtiden; de kan också dra nytta av det etablerade maskineriet och påbörja sina egna riktade samlingar. ”Antalet lingvister som aktivt använder våra standarder och arbetsflöden ökar stadigt. Vi hoppas att den nya versionen av CLICS kommer att bidra till en vidare ökning” säger Simon Greenhill.

Title: The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies

Authors: Christoph Rzymski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalie Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, Johann-Mattis List.

Publication: Scientific Data, DOI: 10.1038/s41597-019-0341-x

Media Contacts:

Christoph Rzymski
Scientific Programmer
Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History
Phone: +49 179-7772995
Email: rzymski@shh.mpg.de

Tiago Tresoldi
Post-Doc, Computer-Assisted Language Comparison (CALC)
Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History
Phone: +49 3641 686-853
Email: tresoldi@shh.mpg.de

Max-Planck-Institut für Menschheitsgeschichte
Max Planck Institute for the Science of Human History

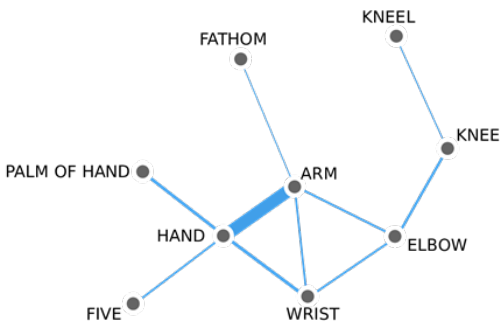
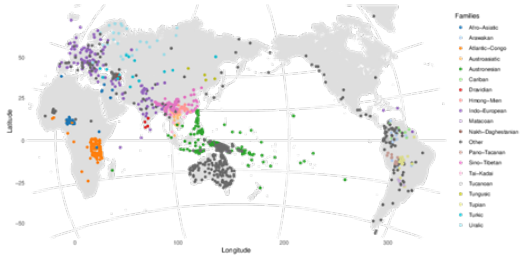
Robert Forkel
Scientific Programmer
Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History
Email: forkel@shh.mpg.de

Johann-Mattis List
Research Group Leader, Computer-Assisted Language Comparison (CALC)
Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History
Phone: +49 1575-2057010
Email: list@shh.mpg.de

AJ Zeilstra / Petra Mader
Public Relations & Press Office
Max Planck Institute for the Science of Human History
Kahlaische Str. 10, 07745 Jena
GERMANY
Phone: +49 (0) 3641 686-950 / 960
Email: presse@shh.mpg.de

Images:

High resolution images are available at:
https://oc.gnz.mpg.de/owncloud/index.php/s/pEFj5qJRNxbz4PZ

Thumbnail	File name, caption and credits
	*File name:* colexification_network.png *Caption:* Colexification network centered on the concepts of “hand” and “arm”. *Photo credit:* J.-M. List, T. Tresoldi
	*File name:* language_map.png *Caption:* Global distribution of languages included in the CLICS3 release, identified by language family. *Photo credit:* S. J. Greenhill