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: La base de datos de asociaciones léxicas interlingüísticas más grande del mundo

La nueva versión de la base de datos define nuevos estándares de investigación reproducible, al ofrecer un enfoque confiable para el estudio cuantitativo de la lingüística

Un grupo de científicos liderado por investigadores del Instituto Max Planck para la Ciencia de la Historia Humana, ha publicado una nueva versión de la Base de datos de colexificaciones interlingüísticas CLICS, que incluye asociaciones léxicas en más de 3100 lenguas. Esta versión ofrece datos léxicos de una magnitud sin precedentes, así como un procedimiento detallado y reproducible para la incorporación de nuevos datos, permitiendo a investigadores de todo el mundo contribuir en versiones futuras.

En toda lengua existen casos en que dos o más conceptos se expresan con la misma palabra, tal como la palabra *vino* en español, que puede referir tanto a una conjugación en pasado del verbo *venir* como a la bebida proveniente de la fermentación de la uva. El comparar los patrones observados de este fenómeno, que los lingüistas llaman *colexificaciones*, en distintas lenguas, proporciona a los investigadores un mayor entendimiento sobre un amplio rango de temas, tales como la percepción humana, la evolución de la lengua y el contacto lingüístico. La tercera entrega de la base de datos CLICS aumenta de manera significativa el número de lenguas, conceptos y fuentes disponibles en versiones anteriores, permitiendo a los investigadores estudiar las colexificaciones a escala global con un detalle y profundidad sin precedentes.

Utilizando procedimientos asistidos por computadora, CLICS facilita la estandarización de bases de datos lingüísticas y ofrece soluciones a muchos de los desafíos usuales en la investigación lingüística. “Mientras que el ingreso de datos ha estado basado mayormente en procedimientos *ad-hoc* en el pasado, nuestros nuevos procedimientos y pautas para una buena práctica, son un paso importante para garantizar la reproducibilidad de la investigación en lingüística”, dice Tiago Tresoldi.

Efectividad de CLICS demostrada en aplicaciones científicas

El poder de CLICS para proveer nueva evidencia al abordar problemas de vanguardia en psicología y cognición ya ha sido ilustrada en un estudio recientemente publicado en Science, que trata la codificación a escala mundial de conceptos relacionados con la emoción. Este estudio compara las redes de colexificación de palabras que denotan emociones en una muestra de lenguas global, revelando que los significados atribuidos a las emociones varían fuertemente entre las distintas familias lingüísticas.

“En este estudio, se utilizó CLICS para estudiar las diferencias en la codificación léxica de las emociones en las lenguas del mundo, pero el potencial de la base de datos no se limita a conceptos relacionados con la emoción. En el futuro se pueden abordar muchas más preguntas interesantes”, dice Johann-Mattis List.

Nuevos estándares y procedimientos permiten la recolección reproducible de datos léxicos globales

Sobre la base de las nuevas pautas para un formato de datos estandarizado en la investigación interlingüística, presentadas por primera vez en 2018 (DOI: 10.1038/sdata.2018.205), el equipo de CLICS logró incrementar la cantidad de datos de 300 lenguas y 1200 conceptos en su base de datos original, a 3156 lenguas y 2906 conceptos en la entrega actual. La nueva versión también garantiza la

reproducibilidad en el proceso de incorporación de datos, lo que contribuye a mejores prácticas en la administración de datos de investigación. “Gracias a los nuevos estándares y procedimientos que hemos desarrollado, nuestros datos no solo son FAIR (“justos”, y también acrónimo en inglés de hallables, accesibles, interoperables y reproducibles), sino que el proceso de conversión de los datos lingüísticos desde sus formas originales a los estándares interlingüísticos es mucho más eficiente que en el pasado”, dice Robert Forkel.

La efectividad del procedimiento desarrollado por CLICS ha sido testeada y confirmada en varios experimentos de validación que involucran a un gran rango de investigadores y estudiantes. Se han realizado dos tareas para estudiantes que resultaron en la creación de nuevas bases de datos y la mejora progresiva de los datos existentes. Se ha enseñado a los estudiantes a seguir todos los pasos de la creación de datos descrita en el estudio, tales como la extracción de datos, su contrastación con los catálogos de referencia y la indentificación de sus fuentes. “Tener gente externa al equipo central que utilice y ponga a prueba nuestras herramientas es esencial y ayuda de manera importantísima a la puesta a punto de todos nuestros procesos”, dice Christoph Rzymski.

Con CLICS y sus procedimientos accesibles a un público más amplio, los investigadores pueden no solo contribuir directamente a la base de datos en el futuro, sino que también pueden beneficiarse de la estructura ya existente y crear sus propias colecciones específicas. “La cantidad de lingüistas que usan nuestros estándares y procedimientos activamente está en constante crecimiento. Esperamos que la entrega de esta nueva versión de CLICS la impulse aun más”, dice Simon Greenhill.

Título: The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies (La base de datos de colexificaciones interlingüísticas, análisis reproducible de polisemias interlingüísticas)

Autores: 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.

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

Contacto con los medios:

Christoph Rzymski
Programador científico
Departamento de Evolución Lingüística y Cultural del Instituto Max Planck para la Ciencia de la Historia Humana
Teléfono: +49 179-7772995
Email: rzymski@shh.mpg.de

Tiago Tresoldi
Investigador postdoctoral en el grupo de Comparación Lingüística Asistida por Computadora (CALC)
Departamento de Evolución Lingüística y Cultural del Instituto Max Planck para la Ciencia de la Historia Humana
Teléfono: +49 3641 686-853
Email: tresoldi@shh.mpg.de

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

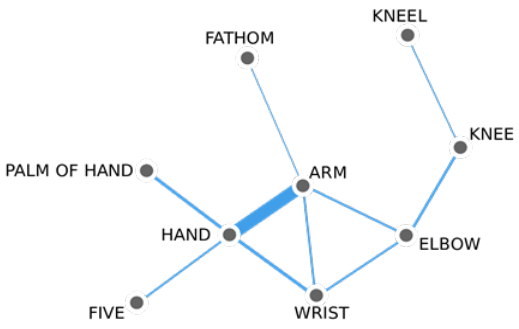
Robert Forkel
Programador científico
Departamento de Evolución Lingüística y Cultural del Instituto Max Planck para la Ciencia de la Historia Humana
Email: forkel@shh.mpg.de

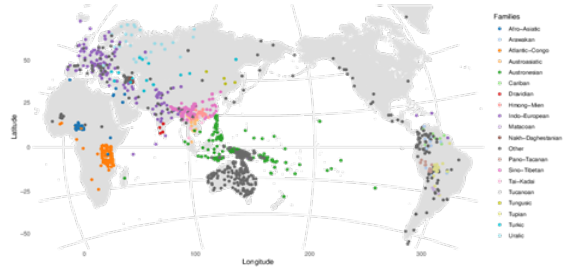
Johann-Mattis List
Líder del grupo de investigación de Comparación Lingüística Asistida por Computadora (CALC)
Departamento de Evolución Lingüística y Cultural del Instituto Max Planck para la Ciencia de la Historia Humana
Teléfono: +49 1575-2057010
Email: list@shh.mpg.de

AJ Zeilstra / Petra Mader
Relaciones Públicas y Oficina de Prensa
Instituto Max Planck para la Ciencia de la Historia Humana
Kahlaische Str. 10
07745 Jena
ALEMANIA
Teléfono: +49 (0) 3641 686-950 / 960
Email: presse@shh.mpg.de

Imágenes:

Imágenes en alta resulución disponibles en:
https://oc.gnz.mpg.de/owncloud/index.php/s/pEFj5qJRNxbz4PZ

Miniatura	**Nombre del archivo, epígrafe y créditos**
	*Nombre:* colexification_network.png *Epígrafe:* Red de colexificaciones alrededor de los conceptos “mano” y “brazo” (“hand” y “arm” respectivamente). *Crédito:* J.-M. List, T. Tresoldi



Nombre: language_map.png

Epígrafe: Distribución global de lenguas incluidas en la base de datos CLICS3, identificando sus familias lingüísticas de pertenencia.

Crédito: S. J. Greenhill