



CLICS: il più grande database di associazioni lessicali trasversali mai pubblicato

La più recente versione del database stabilisce un nuovo standard per la ricerca riproducibile, consentendo un approccio affidabile agli studi linguistici quantitativi

Un team di scienziati, guidato da ricercatori dell'Institut Max Planck per la Scienza della Storia Umana, ha appena pubblicato una nuova versione del [Database di Colessificazioni Linguistiche Trasversali](#) (Database of Cross-Linguistic Colexifications, CLICS), che riporta associazioni lessicali in oltre 3100 lingue. La nuova versione fornisce dati lessicali su una scala senza precedenti e introduce flussi di lavoro riproducibili e dettagliati per la raccolta dati, consentendo agli studiosi di tutto il mondo di contribuire alle versioni future.

Ogni lingua ha casi in cui due o più concetti sono espressi dalla stessa parola, come la parola inglese *fly*, che si riferisce sia all'atto di volare che all'insetto ("mosca"). Confrontando nelle varie lingue questi casi, chiamati "colessificazioni" dai linguisti, i ricercatori possono trovare risposte a un'ampia gamma di domande, come sulla cognizione umana, sull'evoluzione linguistica e sul contatto tra lingue. La terza versione del database CLICS aumenta in modo significativo il numero di lingue, concetti e fonti primarie rispetto alle versioni precedenti, consentendo ai ricercatori di studiare colessificazioni su scala globale a un livello di dettaglio e profondità senza precedenti.

Attraverso flussi di lavoro assistiti da computer, CLICS facilita la standardizzazione dei database linguistici e offre soluzioni alle varie sfide che persistono nella ricerca linguistica. "In passato l'integrazione dei dati linguistici era basata su procedure *ad hoc*, ma i nostri nuovi flussi di lavoro e le relative linee guida sono un passo importante per garantire la riproducibilità della ricerca linguistica", afferma Tiago Tresoldi.

Efficacia del database CLICS dimostrata nelle applicazioni di ricerca

Le capacità di CLICS di fornire nuove prove per affrontare problematiche all'avanguardia in psicologia e cognizione è stata dimostrata in [uno studio pubblicato recentemente su Science](#), incentrato sulla codifica di concetti emotivi su scala mondiale. Lo studio, che ha confrontato un campione globale di reti di colessificazione relative a parole usate per esprimere concetti emozionali, ha rivelato che i significati di queste parole emotive variano notevolmente tra le famiglie linguistiche.

"In questo studio, CLICS è stato utilizzato per studiare le differenze nella codifica lessicale delle emozioni nelle lingue di tutto il mondo, ma il potenziale del database non si limita ai concetti emotivi. Molte altre questioni di grandissimo interesse potranno essere affrontate in futuro", afferma Johann-Mattis List.

Nuovi standard e flussi di lavoro consentono la raccolta riproducibile di dati lessicali globali

A partire dalle nuove [linee guida per formati di dati standardizzati per la ricerca linguistica trasversale](#), presentate per la prima volta nel 2018 (DOI: [10.1038/sdata.2018.205](#)), il team di CLICS è stato in grado di aumentare la quantità di dati da 300 lingue e 1200 concetti nel database originale in 3156 lingue e 2906 concetti nella versione appena rilasciata. Questa nuova versione garantisce la riproducibilità del processo di integrazione dei dati di terze parti, in linea con le migliori pratiche nella gestione dei dati per la ricerca scientifica. "Grazie ai nuovi standard e ai flussi di lavoro che abbiamo sviluppato, non solo i nostri dati sono 'equi' (dall'acronimo FAIR in inglese, ossia localizzabili, accessibili, interoperabili e riproducibili), ma il processo di 'sollevamento' dei dati linguistici dai loro formati originali a questi nuovi standard sono ora molto più efficienti di prima", afferma Robert Forkel.



L'efficacia del flusso di lavoro sviluppato per CLICS è stata testata e confermata in vari esperimenti di validazione che hanno coinvolto una vasta gamma di ricercatori e studenti. Sono stati sviluppati due diversi compiti di ricerca, che hanno portato alla creazione di nuovi set di dati e al progressivo miglioramento di quelli esistenti. In entrambi gli esperimenti, gli studenti universitari hanno avuto il compito di lavorare con le diverse fasi della creazione di database descritti nello studio, come l'estrazione di dati grezzi, la mappatura degli input con cataloghi di riferimento e l'identificazione di riferimenti bibliografici. "Avere persone al di fuori del core team che utilizzano e collaudano i nostri strumenti è essenziale e di grande aiuto nei processi di messa a punto", afferma Christoph Rzymiski.

Con CLICS e il suo flusso di lavoro disponibili per un pubblico più ampio, gli studiosi possono non solo contribuire direttamente alle future espansioni del database ma possono anche trarre profitto della struttura stabilita per avviare raccolte personalizzate. "Il numero di linguisti che utilizzano attivamente i nostri standard e flussi di lavoro è in costante aumento. Speriamo che il rilascio di questa nuova versione di CLICS li propaghi ulteriormente", afferma Simon Greenhill.

Titolo: Il Database di Colessificazioni Linguistiche Trasversali, analisi riproducibile delle polisemie interlinguistiche

Autori: Christoph Rzymiski, 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.

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

Contatti per i media:

Christoph Rzymiski
Programmatore scientifico
Dipartimento di Evoluzione Linguistica e Culturale dell'Istituto Max Planck per la Scienza della Storia Umana
Telefono: +49 179-7772995
E-mail: rzymiski@shh.mpg.de

Tiago Tresoldi
Ricercatore Post-Doc, Comparazione Linguistica Assistita da Computer (CALC)
Dipartimento di Evoluzione Linguistica e Culturale dell'Istituto Max Planck per la Scienza della Storia Umana
Telefono: +49 3641 686-853
E-mail: tresoldi@shh.mpg.de

Robert Forkel
Programmatore scientifico
Dipartimento di Evoluzione Linguistica e Culturale dell'Istituto Max Planck per la Scienza della Storia Umana
E-mail: forkel@shh.mpg.de

Johann-Mattis List
Leader del gruppo di ricerca, Comparazione Linguistica Assistita da Computer (CALC)



Dipartimento di Evoluzione Linguistica e Culturale dell'Istituto Max Planck per la Scienza della Storia Umana

Telefono: +49 1575-2057010

E-mail: list@shh.mpg.de

AJ Zeilstra / Petra Mader

Pubbliche relazioni e ufficio stampa

Istituto Max Planck per la Scienza della Storia Umana

Kahlaische Str. 10

07745 Jena

GERMANIA

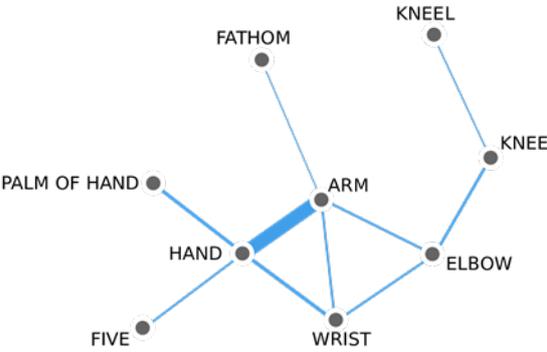
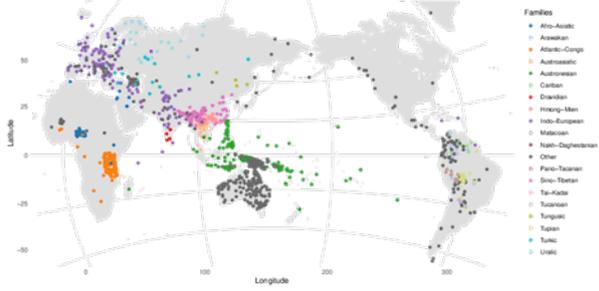
Telefono: +49 (0) 3641 686-950 / 960

E-mail: presse@shh.mpg.de

Immagini:

Le immagini ad alta risoluzione sono disponibili su:

<https://oc.gnz.mpg.de/owncloud/index.php/s/pEFj5qJRNxbz4PZ>

Miniatura	Nome file, didascalia e crediti
	<p><i>Nome file:</i> colexification_network.png</p> <p><i>Didascalia:</i> Rete di colessificazione centrata sui concetti di "mano" e "braccio".</p> <p><i>Crediti:</i> J.-M. List, T. Tresoldi</p>
	<p><i>Nome file:</i> language_map.png</p> <p><i>Didascalia:</i> Distribuzione globale delle lingue incluse in CLICS3, identificate dalla famiglia linguistica.</p> <p><i>Crediti:</i> S. J. Greenhill</p>