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

CLICS: O maior banco de dados de associações lexicais transversais já publicado

A mais nova versão do banco de dados estabelece um novo padrão para a pesquisa reproduzível, possibilitando uma abordagem confiável para estudos linguísticos quantitativos

Uma equipe de cientistas, liderada por pesquisadores do Instituto Max Planck para a Ciência da História Humana, acaba de publicar uma nova versão do Banco de Dados de Colexificações Linguísticas Transversais ("Database of Cross-Linguistic Colexifications", CLICS), o qual inclui associações lexicais entre mais de 3100 idiomas. A nova versão oferece dados lexicais em uma escala sem precedentes e introduz um fluxo de trabalho reproduzível e detalhado para reunir dados, permitindo que estudiosos de todo o mundo contribuam para futuras versões.

Toda língua possui casos em que dois ou mais conceitos são expressos pela mesma palavra, como a palavra inglesa *fly*, que se refere tanto ao ato de voar quando ao inseto ("mosca"). Ao comparar padrões desses casos, chamados de "colexificações" pelos linguistas, os pesquisadores podem encontrar respostas para uma ampla gama de questões, como de cognição humana, de evolução linguística, e de contato entre idiomas. A terceira versão do banco de dados CLICS aumenta significativamente o número de línguas, conceitos e fontes primárias das versões anteriores, permitindo que pesquisadores estudem colexificações em escala global em um nível de detalhe e profundidade sem precedentes.

Por meio de fluxos de trabalho assistidos por computadores, o CLICS facilita a padronização de bancos de dados linguísticos e fornece soluções para vários desafios que persistem na pesquisa linguística. "No passado a integração de dados linguísticos costumava ser baseada em procedimentos *ad hoc*, mas nossos novos fluxos de trabalho e diretrizes são um passo importante para garantir a reprodutibilidade da pesquisa linguística", diz Tiago Tresoldi.

Eficácia do CLICS demonstrada em aplicações de pesquisa

A capacidade do CLICS em fornecer novas evidências para abordar questões de ponta em psicologia e cognição foi recentemente demonstrada em um estudo publicado na revista Science, focado na codificação de conceitos emocionais em escala mundial. O estudo comparou as redes de colexificação de palavras para conceitos emocionais em uma amostra global de idiomas, revelando que os significados das palavras emocionais variam muito entre as famílias linguísticas.

"Neste estudo, o CLICS foi usado para estudar diferenças na codificação lexical de emoções em idiomas de todo o mundo, mas o potencial do banco de dados não se limita aos conceitos de emoção. Muitas outras questões de grande interesse poderão ser abordadas no futuro," relata Johann-Mattis List.

Novos padrões e fluxos de trabalho permitem uma coleta reproduzível de dados lexicais em escala global

Com base nas novas diretrizes para formatos de dados padronizados destinados à pesquisa linguística transversal, apresentadas em 2018 (DOI: 10.1038/sdata.2018.205), a equipe do CLICS conseguiu expandir a quantidade de dados de 300 idiomas e 1200 conceitos na versão original para 3156 idiomas e 2906 conceitos na atual. A nova versão também garante a reprodutibilidade do processo de integração de dados terceiros, em linha com as melhores práticas em gerenciamento de dados de pesquisa científica. "Graças aos novos padrões e fluxos de trabalho que desenvolvemos, nossos dados não são apenas 'justos' (do acrônimo inglês FAIR, para localizáveis, acessíveis, interoperáveis e reproduzíveis), mas o próprio processo de elevação de dados linguísticos de seus formatos brutos para esses novos padrões também é agora muito mais eficiente do que antes", diz Robert Forkel.

A eficácia do fluxo de trabalho desenvolvido para o CLICS for testada e confirmada em diferentes experimentos de validação envolvendo uma grande variedade de pesquisadores e estudantes. Duas diferentes tarefas de pesquisa foram desenvolvidas, resultando na criação de novos conjuntos de dados e na melhoria progressiva dos dados existentes. Nessas tarefas, estudantes de graduação foram incumbidos de trabalhar com as diferentes etapas da criação de bancos de dados descritas no estudo, como a extração de dados brutos, o mapeamento de entrada com catálogos de referência e a identificação das referências bibliográficas. “Ter pessoas de fora da equipe principal usando e testando suas ferramentas é essencial e de extrema importância no ajuste fino dos processos”, afirma Christoph Rzymski.

Com o CLICS e seu fluxo de trabalho disponível a um público mais amplo, os estudiosos não apenas podem contribuir diretamente com futuras expansões do banco de dados: eles também podem se beneficiar com a estrutura estabelecida e iniciar suas próprias coleções sob medida. “O número de linguistas que usam ativamente nossos padrões e fluxos de trabalho está em constante aumento. Esperamos que o lançamento desta nova versão do CLICS seja capaz de propagá-los ainda mais”, completa Simon Greenhill.

Título: O Banco de Dados de Colexificações Linguísticas Transversais para a análise reproduzível de polissemias linguísticas cruzadas

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.

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

Contatos para mídia:

Christoph Rzymski
Programador Científico
Departamento de Evolução Linguística e Cultural do Instituto Max Planck para a Ciência da História Humana
Telefone: +49 179-7772995
E-mail: rzymski@shh.mpg.de

Tiago Tresoldi
Pós-doutor, Comparação de Idiomas Assistida por Computadores (CALC)
Departamento de Evolução Linguística e Cultural do Instituto Max Planck para a Ciência da História Humana
Telefone: +49 3641 686-853
E-mail: tresoldi@shh.mpg.de

Robert Forkel
Programador Científico
Departamento de Evolução Linguística e Cultural do Instituto Max Planck para a Ciência da História Humana
E-mail: forkel@shh.mpg.de

Johann-Mattis List
Líder de Grupo de Pesquisa, Comparaćão de Idiomas Assistida por Computadores (CALC)

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

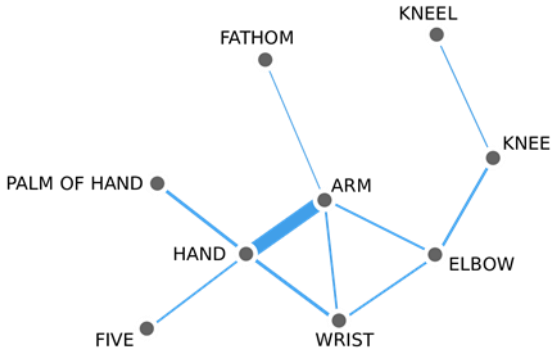
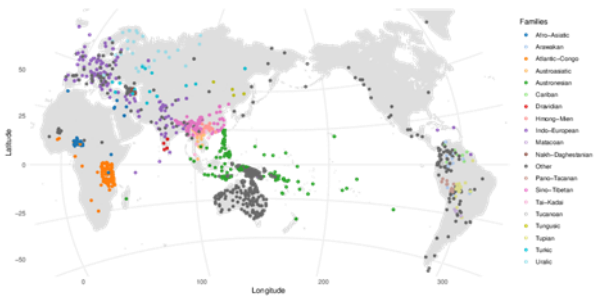
MAX-PLANCK-GESELLSCHAFT

Departamento de Evolução Linguística e Cultural do Instituto Max Planck para a Ciência da História Humana
Telefone: +49 1575-2057010
E-mail: list@shh.mpg.de

AJ Zeilstra / Petra Mader
Relações Públicas e Assessoria de Imprensa
Instituto Max Planck para a Ciência da História Humana
Kahlaische Str. 10
07745 Jena
ALEMANHA
Telefone: +49 (0) 3641 686-950 / 960
E-mail: presse@shh.mpg.de

Imagens:

Imagens de alta resolução estão disponíveis em:
https://oc.gnz.mpg.de/owncloud/index.php/s/pEFj5qJRNxbz4PZ

Miniatura	**Nome de arquivo, legenda e créditos**
	*Nome de arquivo:* colexification_network.png *Legenda:* Rede de colexificação centrada nos conceitos de “mão” e “braço”. *Créditos:* J.-M. List, T. Tresoldi
	*Nome de arquivo:* language_map.png *Legenda:* Distribuição global de idiomas incluídos no CLICS3, identificados por família de idiomas. *Créditos:* S. J. Greenhill