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

CLICS：世界上最大的跨语言词汇关联数据库

最新版本的数据库为研究的可再现性设定了新标准，为语言学的定量研究提供了可靠的方法

由马克斯·普朗克人类历史科学研究所的学者领导的一组科学家发布了最新版本的跨语言共词化数据库（Database of Cross-Linguistic Colexifications CLICS），涵盖了超过 3100 种语言的关联词汇。新版本的数据库以前所未有的规模提供了词汇数据，并为数据聚合提供了详细、可复制的工作流，使世界各地的学者都可以为数据库的未来版本做出贡献。

在每种语言中，都有两个或两个以上的概念用同一词表示的情况，例如英语单词 fly，既指飞行的行为，又指苍蝇。语言学家将这种模式称为共词化（colexification）。通过比较不同语言中的共词化现象，研究人员可以洞悉广泛的问题，包括人类的感知，语言的演变和语言的接触。CLICS 数据库的第三部分显著增加了早期版本中可用的语言、概念和数据源的数量，从而使研究人员能够以前所未有的细节和深度在全球范围内研究共词化现象。

借助详细的计算机辅助工作流程，CLICS 促进了语言数据集的标准化，并为语言研究中的许多持续挑战提供了解决方案。Tiago Tresoldi 表示：“过去数据汇总通常是通过临时决定的步骤完成，但我们的新工作流程和最佳实践指南是确保语言研究可重复性的重要一步。”

研究应用证明 CLICS 的有效性

最近在《科学》杂志上发表的一项研究中已经说明了 CLICS 提供新证据以解决心理学和认知方面的前沿问题的能力，该研究集中在情感概念在全球语言中的不同表达。这项研究比较了来自全球语言样本中的用于情感概念的共词化网络，并发现情感的含义在不同的语言家族中差异很大。

“在这项研究中，使用 CLICS 来研究世界各地语言在情感词汇方面的差异，但是数据库的潜力并不局限于情感概念。”Johann-Mattis List 说，“我们将来还会解决更多有趣的问题。”

新标准和工作流程为收集可复制的全球词汇数据提供可能性

基于 2018 年首次提出的跨语言研究中标准化数据格式的新指南（guidelines for standardized data formats in cross-linguistic research，DOI：10.1038 / sdata.2018.205），CLICS 团队把数据库从 300 种语言和 1200 种概念增加到了 3156 种语言和 2906 种概念。新版本还保证了数据聚合（data aggregation）过程的可重复性，符合研究数据管理中的最佳实践原则。 “由于我们开发了新的标准和工作流程，我们的数据不仅是公开、公平的（可查找、可访问、可互操作和可再现），而且将语言数据从其原始形式提升到我们的跨语言标准的过程也更加高效。” Robert Forkel 说。

为 CLICS 开发的工作流的有效性已经在涉及大量学者和学生的各种验证实验中得到测试和证实。两项不同的学生任务为此展开，创建了新的数据集并逐步改进了现有数据。这两项任务要求学生完成研究中描述的数据集，并创建的不同步骤，例如数据提取，数据映射（到参考目录）和源识别。“让核心团队以外的人使用和测试你的工具是必不可少的，这对微调所有流程有很大帮助，” Christoph Rzymski 说。

随着 CLICS 及其工作流程可供更广泛的受众使用，学者们将来不仅可以直接对数据库做出贡献，还可以在数据库中使用。他们还可以从既有的设备中获利并开始自己的目标收藏。 “积极使用我们的标准和工作流程的语言学家的数量正在不断增加。我们希望这个新版本的 CLICS 能够进一步传播它们。” Simon Greenhill 说。

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

题目: 跨语言共词化数据库，可复制的跨语言多义词汇分析

作者: 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.

发表刊物: Scientific Data, DOI:

联系方式:

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

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

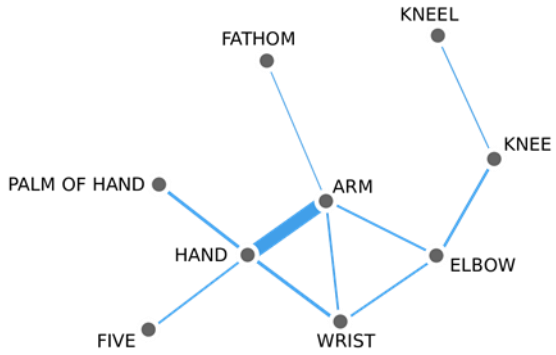
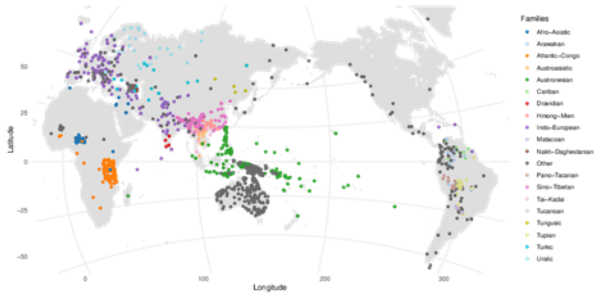
Johann-Mattis List
Research Group Leader, Computer-Assisted Language Comparison (CALC)
Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History
Phone: +49 1575-2057010
Email: list@shh.mpg.de

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:

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

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:* “手”和“手臂”的共词关系网 *图片来源:* J.-M. List, T. Tresoldi
	*File name:* language_map.png *Caption:* CLIC3 中所包含的语系在全球范围内的分布 *图片来源:* S. J. Greenhill