Max-Planck-Institut für Menschheitsgeschichte

Max Planck Institute for the Science of Human History

MAX-PLANCK-GESELLSCHAFT

# CLICS: Крупнейшая база данных лексической полисемии в языках мира

**Новая версия базы данных задает новые стандарты в сфере воспроизводимости результатов научных работ и предлагает новый подход к квантитативной лингвистике**

**Команда ученых, возглавляемая исследователями Института истории человечества Общества Макса Планка, выпустила новую версию базы данных кросс-лингвистических колексификаций (CLICS), представленных в более чем 3100 языках. Последняя версия представляет собой крупнейшую лексическую базу данных с подробной и воспроизводимой схемой для агрегирования данных, которая позволит ученым со всего мира вносить вклад в дальнейшее развитие CLICS.**

В каждом языке есть понятия, выраженные одним и тем же словом. Например, в английском языке *fly* относится как к насекомому, так и к процессу полета. Сравнительный анализ этого лингвистического явления в разных языках, известного как «колексификация», позволит ученым получить лучшее представление о человеческом восприятии, развитии языка, языковом контакте и многом другом. В третьем выпуске базы данных собрано значительно бóльшее количество языков, понятий и источников данных, чем в предыдущих версиях, что обеспечит возможность изучения колексефикаций в глобальном масштабе и с невероятной точностью.

Благодаря компьютеризации рабочих процессов, CLICS способствует стандартизации лингвистических данных и предлагает способы решения многих проблем в сфере лингвистических исследований. «Если раньше агрегирование данных осуществлялось по необходимости, то сейчас разработанная нами схема рабочего процесса и руководство по передовой практике гарантируют воспроизводимость лингвистических исследований», – сообщил Тьяго Тресольди.

## Эффективность применения CLICS в исследовательской сфере

Возможности CLICS по предоставлению новых данных для изучения современных проблем в области психологии и человеческого познания были продемонстрированы в недавнем исследовании в журнале Science. В результате сравнения систем колексифкации среди слов для обозначения эмоций ученые обнаружили, что в зависимости от языковой семьи значения этих слов сильно различаются.

«В данном исследовании база данных CLICS использовалась для изучения лексических способов выражения эмоций, однако этим ее потенциал не исчерпывается. В дальнейшем, она сможет помочь ответить на многие другие вопросы», – заявил Йоханн-Маттис Лист.

## Новые стандарты и практики обеспечат воспроизводимость сбора лексических данных

Опираясь на новые принципы по стандартизации данных в кросс-лингвистических исследованиях, опубликованных в 2018 году (DOI: 10.1038/sdata.2018.205), команда разработчиков CLICS увеличила объем данных с 300 языков и 1200 понятий в первоначальном варианте до 3156 языков и 2906 понятий в последней версии. Кроме того, новейшая версия CLICS обеспечивает воспроизводимость процесса сбора и обработки данных в соответствии с передовым опытом в научной сфере управления данными. «Благодаря нашим стандартам и практикам, CLICS не только соответствуют принципам FAIR (с англ. доступность информации для поиска и получения, возможность ее интеграции с другими ресурсами и

воспроизводимость), но и сам процесс трансформации первичных данных в универсальный кросслингвистический формат стал намного эффективнее», – сообщил Роберт Форкель.

Эффективность организации рабочего процесса CLICS была протестирована и подтверждена в различных экспериментах, в которых приняли участие многие ученые и студенты. Результатом двух заданий, порученных студентам, стало создание новых наборов данных, а также постепенное улучшение существующих данных. Задания включали в себя проработку всех этапов подготовки данных, включая их извлечение и сопоставление (со справочными каталогами), а также идентификацию их источников. «Важно, чтобы кто-то не из рабочей группы протестировал разработанные инструменты: такая практика помогает в усовершенствовании рабочих процессов», – сообщил Кристоф Рзимски.

С помощью базы данных CLICS и ее системы рабочих процессов, доступных для широкой аудитории, ученые смогут не только участвовать в ее развитии, но и пользоваться ее преимуществами для сбора данных в собственных целях. «Со временем все больше лингвистов руководствуются нашими принципами и практиками. Мы надеемся, что новая версия CLICS будет и дальше способствовать их популяризации», – поделился Саймон Гринхилл.

**Заголовок:** База данных кросслингвистических колексификаций, воспроизводимый анализ кросслингвистической полисемии

**Авторы:** Кристоф Рзимски, Тьяго Тресольди, Саймон Дж. Гринхилл, Мей-Шин Ву, Натанаэль Э. Швайкард, Мария Коптьевская-Тамм, Фолькер Гаст, Тимотеус А. Бодт, Абби Хангтан, Гереон А. Кайпинг, Софи Чанг, Юнфан Лай, Наталья Морозова, Хайни Арьява, Наталия Хюблер, Изекиль Койл, Стив Пеппер, Марианн Проос, Бриана Ван Эппс, Ингрид Бланко, Каролин Хундт, Сергей Монахов, Кристина Пьяных, Саллона Рамеш, Рассел, Д. Грей, Роберт Форкель, Йоханн-Маттис Лист.

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

**Контактные данные для СМИ:**

Кристоф Рзимски
Программист-исследователь
Отдел лингвистики и культурной эволюции при Институте истории человечества Общества Макса Планка
Тел.: +49 179-7772995
E-mail: rzymski@shh.mpg.de

Тьяго Тресольди
Постдок, компьютеризированное сравнение языков (CALC)
Отдел лингвистики и культурной эволюции при Институте истории человечества Общества Макса Планка
Тел.: +49 3641 686-853
E-mail: tresoldi@shh.mpg.de

Роберт Форкель
Программист-исследователь
Отдел лингвистики и культурной эволюции при Институте истории человечества Общества Макса Планка
E-mail: forkel@shh.mpg.de

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

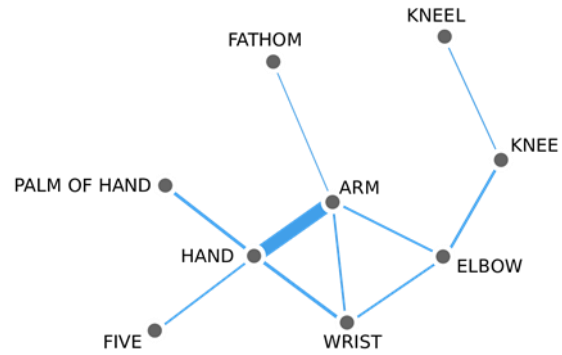
MAX-PLANCK-GESELLSCHAFT

Йоханн-Маттис Лист
Руководитель исследовательской группы, компьютеризированное сравнение языков (CALC)
Отдел лингвистики и культурной эволюции при Институте истории человечества Общества Макса Планка
Тел.: +49 1575-2057010
E-mail: list@shh.mpg.de

ЕйДжей Зейлстра / Петра Мадер
Связи с общественностью, пресс-служба
Институт истории человечества Общества Макса Планка
Калайше-штрассе 10
07745, г. Йена
ГЕРМАНИЯ
Тел.: +49 (0) 3641 686-950 / 960
E-mail: presse@shh.mpg.de

**Изображения:**

*Изображения в высоком разрешении доступны на*
https://oc.gnz.mpg.de/owncloud/index.php/s/pEFj5qJRNxbz4PZ

| **Изображение в миниатюре** | **Файл, подпись, авторство** |
|---|---|
|  | *Файл*: colexification_network.png<br><br>*Подпись*: Система колексификации, выстроенная вокруг понятий «hand» (рука ниже запястья) и «arm» (выше запястья).<br><br>*Авторы*: Й.-М. Лист, Т. Тресольди |
|  | *Файл*: language_map.png<br><br>*Подпись*: Распределение языков на карте мира, включенных в 3-ю версию CLICS и сгруппированных по генетической принадлежности.<br><br>*Автор*: С. Дж. Гринхилл |