Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

CLICS⁴ as CLLD Web Application

Annika Tjuka, Robert Forkel, and Johann-Mattis List
DLCE / DLCE / Chair for Multilingual Computational Linguistics
Leipzig / Leipzig / University of Passau

The fourth version of the Database of Cross-Linguistic Colexifications (https://clics.clld.org) was published last year. Now, we launched the accompanying web application that presents the data for interactive inspection and exploration. The study presents the application in brief and discusses also the development of cross-linguistic colexification databases in the future.

1 Introduction

It has been more than 10 years ago that the first version of the Database of Cross-Linguistic Colexifications (CLICS) was published (List et al. 2014). Based on the notion of *colexification* by François (2008), a cover term for *polysemy* and *homonymy*, referring to multiple senses expressed by identical word forms in one and the same language, CLICS extracted colexification data from wordlist collections, such as the Intercontinental Dictionary Series (Key and Comrie 2016) and presented the data in the form of a web application that could be interactively browsed by interested users (Mayer et al. 2014). With the publication of CLICS⁴ (Tjuka et al. 2025), the database has had four big installments now, after major updates were made in 2018 (List et al. 2018) and 2020 (Rzymski et al. 2020). CLICS⁴ was initially only published as a dataset, without an accompanying web application. This application has now been finalized in the form of a CLLD application (Forkel 2014) that replaces the previous CLICS³ database (https://clics.clld.org).

Computer-Assisted Language Comparison in Practice
Volume 9, Number 1
URL: https://calc.hypotheses.org/9261 DOI: 10.15475/calcip.2026.1.4

Published on 29/04/2026

2 Background

The first version of CLICS was mainly based on a single dataset – the Intercontinental Dictionary Series, later published in Key and Comrie (2016). Since CLICS², CLICS has been based on the aggregation of several datasets which were unified by converting the original data into the formats suggested by the Cross-Linguistic Data Formats (CLDF) initiative (Forkel et al. 2018). This meant specifically that the translational equivalents that were used in the questionnaires of individual comparative wordlists were mapped to Concepticon concept sets (List et al. 2025), in order to identify concepts recurring across different datasets, and that the language varieties were linked to Glottolog (Hammarström et al. 2026) in order to obtain standardized access to additional metadata on individual languages (their families, geographic coordinates, etc.). While CLICS³ enhanced the aggregation procedure from individual CLDF representations of individual datasets, the fourth installment of CLICS (Tjuka et al. 2025) introduced not only new datasets, but also several novel ideas regarding the handling and analysis of colexification data.

These innvotations include (1) a novel handling of concept relations, by which broad concepts that cover multiple senses, like ARM OR HAND, were represented in *separated form*, allowing to compare them directly with the colexified concepts ARM and HAND in separation; (2) a more rigid selection of *eligible* language varieties and concepts, restricting the selection of language varieties to those exceeding a certain number of basic concepts; (3) the exclusion of datasets where phonetic transcriptions are lacking, following the practice established in the second installment of the Lexibank repository (Blum et al. 2025), and (4) a more principled representation of colexifications in line with the possibilities offered by CLDF.

Taken together, these modifications make CLICS⁴ into a database that crucially differs from previous versions of CLICS. Similar to Lexibank², where we decided to exclude datasets where phonetic transcriptions are lacking or cannot be standardized, CLICS⁴ is now entirely built from data available in phonetic transcriptions. While phonetic transcriptions are not necessarily essential for the computation of colexification data, they may turn out to be crucial for additional colexification studies going beyond the level of *full* colexifications, taking *partial* colexifications in different versions into account (List 2023; Norcliffe and Majid 2024; Tjuka and List 2024; Bocklage et al. 2024; Urban 2011).

Our decision to split certain concepts into two separate concepts can be seen as an important step to cope with the problems introduced by the aggregation procedure underlying CLICS since the publication of its second version. The inspiration for this

modification goes back to the publication of the Lexibank repository, where we treated certain frequently recurring colexifications as features that could be compared across languages (List, Hill, and Forkel 2022). In CLICS² and CLICS³, the concepts ARM and HAND would have been represented by three separate concepts – depending on the underlying dataset –, namely ARM, HAND, and ARM OR HAND. Since datasets in which ARM OR HAND can be found in the questionnaire typically do not have separate entries for ARM and HAND, respectively, this meant that the colexification between ARM and HAND was never measured for those datasets in which only ARM OR HAND occurred as a single concept. By separating these entries as transparently as possible, CLICS⁴ now provides more realistic colexification counts for several concepts that are frequently colexified in the languages of the world.

As a final improvement, the representation of all computed colexifications in CLDF now allows for a much more fine-grained analysis of individual colexification patterns. Thus, while the network representation underlying earlier CLICS versions only provided "positive information" by drawing a link between concepts where colexifications had been observed, the new data representation also allows us to take "negative information" (the number of cases where no colexifications are attested) as well as missing information (the cases where concepts are missing) into account. As a result, we can infer much more detailed information from CLICS⁴ than from its predecessors (Snee and List 2026).

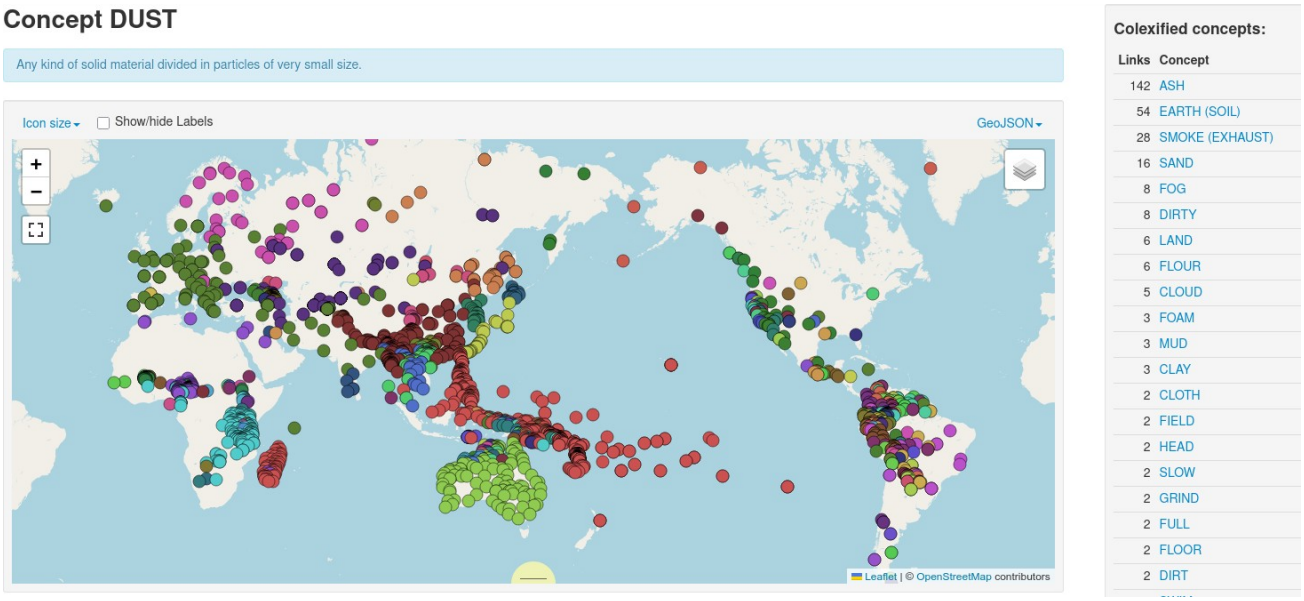

Figure 1: Colexifications involving DUST.

3 CLLD Application

While our initial focus for the creation of CLICS⁴ was the improvement of the data underlying CLICS, we were aware that the CLLD application that had been available since the publication of CLICS² was an integral aspect of the database, given that it allows users to browse the colexifications in an interactive manner. The gap is now closed with the publication of the new CLLD application (Tjuka et al. 2026) that offers users the possibility to query the colexifications for individual concepts, as shown for the colexifications attested for the concept DUST in Figure 1. (see https://clics.clld.org/parameters/2). The individual forms in which a given colexification is attested can also be queried directly through the web application (see https://clics.clld.org/edges/2-646 for the colexification of DUST and ASH).
ASH).

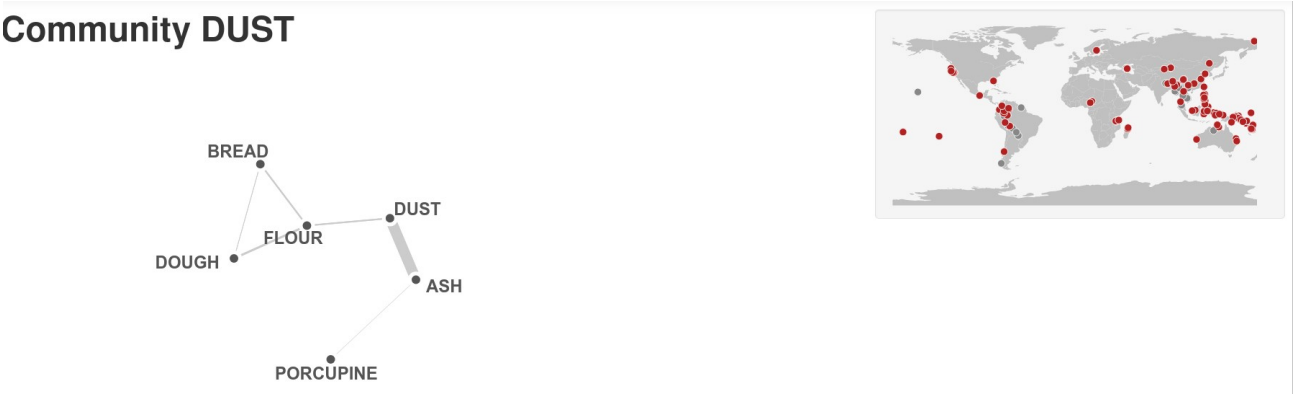

Figure 2: Network representation of a a subgraph of CLICS⁴ with the central concept DUST.

In addition, a network representation in which all concepts involved in an Infomap community (Rosvall and Bergstrom 2008), provided as part of the CLDF representation of the data, allows for the interactive inspection of colexification networks, using a visualization that shows the geographic locations of the languages in which particular colexifications occur (see Mayer et al. 2014 for details), as illustrated in Figure 2.

While these features were also available in earlier versions of CLICS, an advantage of our attempts to increase not only the data in quantity but also improve their quality can be found in the handling of the references that now consistently link to the original datasets, presenting them with unified citations that make it easy to go back to the original sources and check how much they were modified when aggregating them in CLICS⁴. As can be seen from Figure 3, CLICS⁴ now provides not only DOIs to original datasets on Zenodo, but also offers basic statistics, information on the concept lists, and references for the original data.

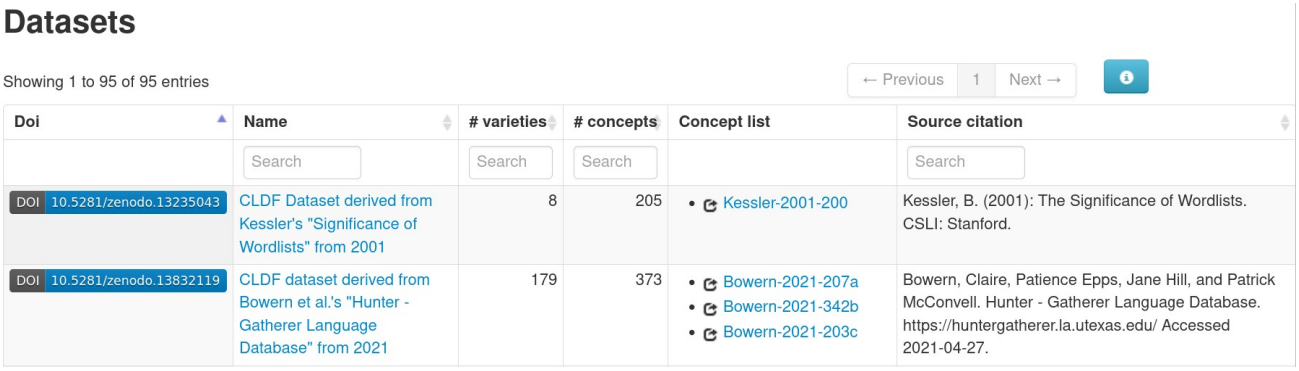

Figure 3: Reference handling of original data in CLICS⁴.

4 Future of CLICS

In the team that was responsible for the launch of CLICS⁴, we have been discussing the future of CLICS for a long time now. We decided that the application in the current form, based on a dataset and a CLLD application that offers a particular view on the data, won't be further updated in the future. Our experience from the past shows that it becomes more and more difficult to manage the numerous parameters that go into the creation of CLICS. Since each parameter we decide for can also be seen as an arbitrary decision that we make, which may conflict with individual research interests of individual users, it seems that it will be more important to reduce further updates of CLICS to some core features. These may include the aggregation of new datasets in a unified CLDF repository, as we have been doing with CLICS⁴, along with the creation of basic code that would allow users to compute colexifications in different forms and inspect them in their own work.

A web application like the CLLD application representing parts of CLICS⁴ now will always offer a reduced view of the original data. With growing degrees of standardization, the possibilities to analyze data in different ways increase further. As a result, it will become more difficult to present the data in a way that reconciles individual use cases. With the introduction of algorithms that compute partial colexifications in various ways (List 2023; Blum et al. 2025), the possibilities of analyzing CLICS data in individual ways increase even further. Instead of investing our efforts into improving the web application, it seems therefore better to improve the ways in which individual analyses of the data can be shared in consistent ways. While we have not found sufficient solutions in this regard, we hope that future analyses will help us to establish good enough practices that allow users with different computational expertise to inspect and use CLICS data in their work.

5 Concluding Remarks

With the publication of CLICS⁴ as a CLLD application, we consider the fourth installment of CLICS as a final step. At the moment, we can't say whether there will be a fifth version., but we see great potential for various novel colexification analyses to emerge in the near future. Given that the complexity of handling colexification data has increased lately, we see a particular challenge in the consistent and transparent handling of analyses applied to colexification data.

References

Blum, Frederic, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, and Johann-Mattis List. 2025. “Lexibank 2: Pre-Computed Features for Large-Scale Lexical Data [version 2; peer review: 3 approved].” *Open Research Europe* 5 (126): 1–24. https://doi.org/10.12688/openreseurope.20216.2.

Bocklage, Katja, Anna Di Natale, Annika Tjuka, and Johann-Mattis List. 2024. “Directional Tendencies in Semantic Change.” *Humanities Commons*. https://doi.org/10.17613/0y0r-f341.

Forkel, Robert. 2014. “The Cross-Linguistic Linked Data Project.” In *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, edited by Christian Chiarcos, John Philip McCrae, Petya Osenova, and Cristina Vertan, 61–66. Reykjavik: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LDL2014%20Proceedings.pdf.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” *Scientific Data* 5 (180205): 1–10. https://doi.org/10.1038/sdata.2018.205.

François, Alexandre. 2008. “Semantic Maps and the Typology of Colexification: Intertwining Polysemous Networks Across Languages.” In *From Polysemy to Semantic Change*, edited by Martine Vanhove, 163–215. Amsterdam: Benjamins.

Hammarström, Harald, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2026. *Glottolog [Dataset, Version 5.3]*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Key, Mary Ritchie, and Bernard Comrie. 2016. *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

List, Johann-Mattis. 2023. “Inference of Partial Colexifications from Multilingual Wordlists.” *Frontiers in Psychology* 14 (1156540): 1–10. https://doi.org/10.3389/fpsyg.2023.1156540.

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. “CLICS². An Improved Database of Cross-Linguistic Colexifications Assembling Lexical Data with Help of Cross-Linguistic Data Formats.” *Linguistic Typology* 22 (2): 277–306. https://doi.org/10.1515/lingty-2018-0010.

List, Johann-Mattis, Nathan W. Hill, and Robert Forkel. 2022. “A New Framework for Fast Automated Phonological Reconstruction Using Trimmed Alignments and Sound Correspondence Patterns.” In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 89–96. Dublin: Association for Computational Linguistics. https://aclanthology.org/2022.lchange-1.9

List, Johann-Mattis, Thomas Mayer, Anselm Terhalle, and Matthias Urban. 2014. *CLICS: Database of Cross-Linguistic Colexifications. Version 1.0* (version 1.0.0). Marburg: Forschungszentrum Deutscher Sprachatlas. http://clics.lingpy.org.

List, Johann-Mattis, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel. 2025. *CLLD Concepticon [Dataset, Version 3.4.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org/.

Mayer, Thomas, Johann-Mattis List, Anselm Terhalle, and Matthias Urban. 2014. “An Interactive Visualization of Cross-Linguistic Colexification Patterns.” In *Visualization as Added Value in the Development, Use and Evaluation of*

Linguistic Resources. Workshop Organized as Part of the International Conference on Language Resources and Evaluation, 1–8. https://lingulist.de/documents/mayer-et-al-2014-clics-visualization.pdf.

Norcliffe, Elisabeth, and Asifa Majid. 2024. "Word Formation Patterns in the Perception Domain: A Typological Study of Cross-Modal Semantic Associations." *Linguistic Typology* 28 (3): 419–59. https://doi.org/10.1515/lingty-2023-0038.

Rosvall, M., and C. T. Bergstrom. 2008. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proc. Natl. Acad. Sci. U.S.A.* 105 (4): 1118–23.

Rzymski, Christoph, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. "The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross-Linguistic Polysemies." *Scientific Data* 7 (13): 1–12. https://doi.org/10.1038/s41597-019-0341-x.

Snee, David, and Johann-Mattis List. 2026. "Computing Detailed Colexifications with Missing Data Information from the CLICS⁴ Collection." *Computer-Assisted Language Comparison in Practice* 9 (1): 7–18. https://doi.org/10.15475/calcip.2026.1.2.

Tjuka, Annika, Robert Forkel, Christoph Rzymski, and Johann-Mattis List. 2025. "Advancing the Database of Cross-Linguistic Colexifications with New Workflows and Data." In *Proceedings of the 16th International Conference on Computational Semantics*, 1–15. https://aclanthology.org/2025.iwcs-main.1.

———. 2026. *CLICS: Database of Cross-Linguistic Colexifications [Database, Version 4.0]*. Passau: MCL Chair at the University of Passau. https://clics.clld.org/.

Tjuka, Annika, and Johann-Mattis List. 2024. "Partial Colexifications Reveal Directional Tendencies in Object Naming." *Yearbook of the German Cognitive Linguistics Association* 12 (1): 95–114. https://doi.org/10.1515/gcla-2024-0005.

Urban, Matthias. 2011. "Asymmetries in Overt Marking and Directionality in Semantic Change." *Journal of Historical Linguistics* 1 (1): 3–47.

Funding Information

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101044282). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.