Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

How to review concept lists in collaboration (How to do X in linguistics 6)

Annika Tjuka
Department of Linguistic and Cultural Evolution
Max Planck Institute for the Science of Human History

In 2016, the Concepticon was introduced as a reference catalog for linguistic data of various kinds (List et al. 2016). The aim of the Concepticon project is to collect concept lists and link the glosses in the lists to unified concept sets (List et al. 2016; List et al. 2020). The project is a collaborative effort and the group of editors is constantly adding new data to the Concepticon (https://concepticon.clld.org/). In 2019, we implemented a review process that works similar to a submission process at an academic journal and is improving the quality of the resource in many ways. This blog post describes our effort to improve data validity by reviewing every single concept list that is added to the Concepticon. The database is curated openly on GitHub, so you can follow our review process by looking at several examples here: https://github.com/concepticon/concepticon-data

Preparation

The first step is to identify a concept list that should be added to Concepticon. We usually create an issue in the GitHub repository for concept lists that are of interest and work through them as we go. The next step is to prepare the concept list in a way that the glosses can be mapped to the Concepticon concept sets. We published several guidelines for the mapping procedure on this blog and the relevant posts are listed under "Concepticon Guidelines" below (Tresoldi 2019a, 2019b; Tjuka 2020; List 2021). Recently, we also started to add more complex datasets, for example, the Multi-SimLex list (Vulić et al. 2020). Multi-SimLex is a multi-lingual concept list with word pairs that were rated on their semantic similarity. A complex list like this requires an advanced workflow, as it does not follow the standard linear structure of our usual concept lists, which have one word per line. List (2021) described the procedure of how the list was

Computer-Assisted Language Comparison in Practice
Volume 4, Number 3
https://calc.hypotheses.org/2680/

Published on 15/03/2021

linked to Concepticon in a recent blog post. The preparation of the lists often already reveals inaccuracies and can be seen as an additional check on the validity of the data.

A collaborative approach to reviewing

More and more data are becoming available across different research fields and are used for multiple purposes. In linguistics, many fieldworkers decide to make their data collections of recently undocumented languages available either in form of online dictionaries (e.g., Haspelmath 2017) or online databases (Bowern et al. 2020). The problem, so far, is that the data stand on their own and are not comparable and interoperable. In the Concepticon, we integrate a variety of concept lists from different sources and make them FAIR: *findable*, *accessible*, *interoperable*, and *reusable* (Wilkinson et al. 2016). The process of mapping concept lists to the concept sets in Concepticon also includes examining the content of each list in detail and checking its consistency. And since several heads are better than one, we established a collaborative review workflow within the Concepticon GitHub repository that is modeled on the process that academic journals employ when reviewing research articles. Each list in Concepticon is thus automatically checked through the build-in tests in pyconcepticon and goes through a human review based on the dual control principle.

GitHub workflow

A new concept list is added to Concepticon via a pull request (PR) in the GitHub repository. The PR already contains a checklist indicating the changes to be made.

```
- add new concept list
- add new metadata
- add new Concepticon concept sets
- add new Concepticon concept relations
- refine existing Concepticon concept set mappings
- refine Concepticon glosses
- refine Concepticon concept relations
- refine Concepticon concept definitions
- retire data
```

The person who created the PR then selects a moderator from the list of Concepticon editors (List et al. 2020). The Concepticon editors are a group of linguists with expertise in historical linguistics, computational linguistics, and psycholinguistics. The moderator in turn is responsible for the smooth execution of the revision process. Next, the moderator appoints two Concepticon editors as reviewers for the PR. Their task is to check the changes made by the person who submitted the concept list and to make

suggestions for improvements. On the GitHub platform, this process is straightforward because it has a built-in review/comment/approve function that can be used by the reviewers. The reviewers are able to comment on individual lines in each file that was added to the PR.

The main discussion about changes is surrounding the concept list. The reviewers check the mapping of each gloss to a certain Concepticon concept set which requires detailed knowledge of the 3,825 concept sets currently included in the Conception. They need to take into account the ontological category of the concept, the relation to other concepts, and its definition. In addition, reviewers may suggest additional mappings or the need for a new concept set. Usually, the review takes a couple of days depending on the availability of the reviewers and the length of the concept list. We had several intense discussions about which gloss to map to which concept set, but in the end, we always decide on the most sensible mapping, or when in doubt, *unmap* the gloss. For example, when the gloss to *plant/put* was mapped to the concept set 998 PUT, the reviewer can propose to unmap it (i.e., delete the mapping) because the concept set refers to the spatial relation. So if none of the concept sets in Concepticon match a given gloss, we leave them unmapped.

Most of the concept lists provide English glosses, but the Concepticon already includes several mappings to other languages, for instance, Spanish, Chinese, and Russian. To include more languages, we are steadily adding concept lists with glosses in other languages, for example, German (see List 2020). As mentioned above, we recently added a multi-lingual concept list (Multi-SimLex, Vulić et al. 2020) to Concepticon with 12 languages, including Russian, Chinese, Welsh, Kiswahili, and others. While adding concept lists of various languages, it is extremely beneficial that our Concepticon editors consist of bi-/multi-linguals and speak the following languages: Chinese, Danish, Dutch, English, French, German, Italian, Portuguese, Russian, Spanish, and Vietnamese. Especially in the case of the Multi-SimLex list, we were able to spot inconsistencies in the translations across the languages Russian, French, Chinese, and Spanish that were not detected by the checks implemented in the creation of the lists (for details on the translation protocol, see Vulić et al. 2020).

The PR is approved when all conversations about proposed changes have been resolved and the reviewers agree that their comments were taken care of. The moderator is responsible for taking a final look at the changes to be made and once all reviewers have approved the PR, it can be merged into the master branch. The new concept list will thus be added to Concepticon and published with the next release.

Advantages of our approach

The main advantage of having a Concepticon as a resource in the first place is that it allows us to have a variety of data stored in one place and make them available to other researchers. In particular, the Concepticon lets us compare and contrast information about concepts from different sources across several languages. The inconsistencies that we find across concept lists also underscore the importance of a test-driven approach to data curation. The Concepticon is a community effort, so a collaborative review process, such as the one outlined in this blog post, enables us to continually improve linguistic data. Since the Concepticon is openly curated on GitHub, every decision made in the review process is transparent. The scope of the Concepticon is constantly broadening so that it is essential to have a team of reviewers that are familiar with the Concepticon structure and the guidelines for the mappings. The review process ensures that the data in Concepticon are consistent and disambiguated. The fact that our editors speak multiple languages also makes it possible for us to cross-check complex datasets like Multi-SimLex and demonstrate the need for human judgment of word meanings and consistency tests in cross-linguistic datasets for NLP tasks. We encourage other researchers to see our process as an example or contribute their concept lists to Concepticon.

Concepticon guidelines

Johann-Mattis List, "Mapping Multi-SimLex to Concepticon," in Computer-Assisted Language Comparison in Practice, 10/03/2021, https://calc.hypotheses.org/2684.

Johann-Mattis List, "How to handle semantic data with tables (How to do X in linguistics 3)," in Computer-Assisted Language Comparison in Practice, 13/01/2021, https://calc.hypotheses.org/2617.

Annika Tjuka, "Adding concept lists to Concepticon: A guide for beginners," in Computer-Assisted Language Comparison in Practice, 29/01/2020, https://calc.hypotheses.org/2225.

Tiago Tresoldi, "Using pyconcepticon to map concept lists (II)," in Computer-Assisted Language Comparison in Practice, 08/04/2019, https://calc.hypotheses.org/1844.

Tiago Tresoldi, "Using pyconcepticon to map concept lists," in Computer-Assisted Language Comparison in Practice, 01/04/2019, https://calc.hypotheses.org/1820.

References

Bowern, C., Epps, P., Hill, J., & Hunley, K. (2020). Hunter-gatherer language database: A collection of lexical, grammatical, and other information about languages spoken by hunter-gatherers and their neighbors. https://huntergatherer.la.utexas.edu/lexical

Haspelmath, M. (2017). Dictionaria: Farewell to linear dictionaries. Diversity Linguistics Comment. https://dlc.hypotheses.org/971

List, J.-M., Rzymski, C., Greenhill, S. J., Schweikhard, N. E., Pianykh, K., Tjuka, A., Wu, M.-S., & Forkel, R. (2020). Concepticon. A resource for the linking of concept lists (Version 2.4.0.). Max Planck Institute for the Science of Human History. https://concepticon.clld.org/

List, J.-M. (2020). Towards a refined wordlist of German in the Intercontinental Dictionary Series. Computer-Assisted Language Comparison in Practice. https://calc.hypotheses.org/2545.

List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon: A resource for the linking of concept lists. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (pp. 2393–2400). European Language Resources Association (ELRA).

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., & Korhonen, A. (2020). Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. Computational Linguistics, 46(4), 1–51. https://doi.org/10.1162/coli_a_00391

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 1–9. https://doi.org/10.1038/sdata.2016.18