Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Creating a CLDF Wordlist from Heath et al.'s Dogon Comparative Wordlist

Abbie Hantgan[1] and Johann-Mattis List[23]
LLACAN, UMR 8135[1], DLCE[2], Chair of Multilingual Computational Linguistics[3]
CNRS, INALCO, Paris[1], MPI-EVA[2], University of Passau[3]

The Dogon and Bangime Linguistics project (https://dogonlanguages.org) offers a large comparative spreadsheet in which translational equivalents for a huge number of concepts are translated into various Dogon languages. Due to its enormous size, no attempts have been made so far to integrate the spreadsheet with the lexical resources that were compiled as part of the CLDF initiative in order to populate the Lexibank repository. Here, we report a first attempt to circumvent the problems resulting from the size of the spreadsheet by convert not all but parts of the spreadsheet to CLDF Wordlist standards, which allows us to integrate parts of the data with other resources in Lexibank.

1 Introduction

One core resource of the Dogon and Bangime Linguistics project (Heath et al. 2016, https://dogonlanguages.org) is the Dogon Comparative Wordlist, a spreadsheet that lists more than 9000 lexical glosses translated into several Dogon language varieties, spoken in Central-Eastern Mali and Burkina Faso. The spreadsheet was created as the result of intensive field work. In size, it reminds more of a multilingual dictionary than a classical comparative wordlists with only a couple of hundred entries. Due to the internal complexity of the comparative wordlists and its very detailed glosses denoting very finegrained semantic differences that would often only find translational equivalents in a few of the more than two dozen language varieties documented by the Dogon and Bangime Linguistics projects, the underlying concept list has so far not yet been linked with the Concepticon resources (List et al. 2023, https://concepticon.clld.org) and the data has only been integrated with the Lexibank repository through dedicated studies in which words were hand-picked and manually corrected (Hantgan and List 2022, Hantgan et al. 2022).

Computer-Assisted Language Comparison in Practice
Volume 6, Number 2, 59-63
URL: https://calc.hypotheses.org/6329 DOI: 10.15475/calcip.2023.2.2

Published on 16/08/2023

2 Background

Given the importance of information on Dogon languages for the investigation of the origin of the language isolate Bangime (Hantgan 2022), we have long since been trying to convert the comparative wordlist into a proper CLDF wordlist that would allow for an easy integration with additional lexical resources on languages in the region. However, all attempts have failed so far, due ot the sheer size of the wordlist.

The data on Dogon languages shared by the Dogon and Bangime Linguistics project by Heath et al. (2016) currently covers lexical data for 30 language varieties, including several Dogon varieties and the isolate Bangime. The spreadsheet lists languages in separate columns and concepts in separate rows, showing a total of more than 43 columns (with specifications of lexical glosses in English and French and at different levels of detail) and 9667 rows, corresponding to a total of 121,223 word forms across all languages.

3 CLDF Conversion

The crucial strategy for the conversion to CLDF was to work with a reduced selection of entries. This would of course result in a CLDF wordlist much smaller in size than the original comparative wordlist, but it would allow us to pursue the targeted normalization steps that are so crucial for the conversion of data to CLDF for the inclusion in the Lexibank repository (List et al. 2022). Thanks to the flexibility of CLDFBench (Forkel and List 2020), one is never required to convert all data in a given source into cross-linguistic data formats. Thanks to the smaller selection of lexical entries, it was also possible to consistently stanardize the transcriptions of all lexical entries, an enterprise which would have not been very difficult to be carried out for all 121,223 word forms in the data.

3.1 Language Selection

Of the 30 varieties in the original comparative wordlist, 25 Dogon varieties with the largest coverage were selected. Coverage is specifically important for computational and quantitative analyses, since languages with low coverage tend to cause numerous problems when it comes to the application of computational techniques, such as automated cognate detection (List et al. 2018) or phylogenetic analysis (Sagart et al. 2019). The table below shows all languages which were selected for inclusion in the standardized wordlist.

Number	Name	Glottocode
1	Ampari	ampa1238
2	Bankan Tey	bent1238
3	Ben Tey	bent1238
4	Bunoge	buno1241
5	Dogul Dom (Bendiely & Kundialang, BC)	dogu1235
6	Dogul Dom (Kundialang)	dogu1235
7	Donno So	donn1239
8	Gourou	guru1265
9	Jamsay (Douentza area, JH)	
10	Jamsay Mondoro	jams1239
11	Mombo	momb1254
12	Najamba	bond1248
13	Nanga	nang1261
14	Penange	pena1270
15	Perge Tegu	perg1234
16	Tebul Ure	tebu1239
17	Tiranige	tira1258
18	Togo-Kan	togo1254
19	Tommo-So (Tongo Tongo, LM)	
20	Tomo Kan Diangassagou	tomo1243
21	Tomo Kan Segue	tomo1243
22	Toro Tegu	toro1253
23	Yanda Dom	yand1257
24	Yorno So	yorn1234

Table 1: Languages selected for the CLDF wordlist.

3.2 Concept Selection

With its more than 9000 distinct glosses for lexical concepts, the spreadsheets shows much more resemblance to a multilingual dictionary than to a comparative wordlist. For the standardization of lexical data in the form of a CLDF Wordlist, concepts need to be linked to the Concepticon project, but concept lists linked to Concepticon rarely exceed more than 1000 concepts in size. In order to ease the linking process and to avoid a biased selection of an only small number of concepts easy to link, we decided to use a new strategy for the linking process. In a first stage, we carried out an automated linking of the lexical glosses to Concepticon, using the PySem library (https://pypi.org/projects/pysem, List 2021). In a second step, all entries were checked, keeping only those that showed a clear mapping to the Concepticon project. This yielded a concept list of 1811 entries. In a third step, this list was further filtered, by retaining only those concept sets, which also occur in the concept list underlying the Intercontinental Dictionary Series (https://ids.clld.org, Key and Comrie 2016). This helped us to further reduce the size of the concepts to 944 items.

While this approach to concept selection was carried out in a semi-automated way very specific to the Dogon data, we think that the general principle of subsampling smaller amounts of data from large lists of lexical glosses may be useful for many future applications, specifically when working, for example, with dictionary data, from which one wishes to subset a certain amount of concepts in order to convert dictionaries into a wordlist.

3.3 Phonetic Transcription

We used the standard procedure of creating orthography profiles (Moran and Cysouw 2018 with the help of the Lexibank pipeline (List et al. 2022). Creating the orthography profile proved particularly challenging, given that tonal information was in part annotated on vowels -- which makes it difficult to account for tone in subsequent automated analyses on shared cognates -- and that word forms often cointained unstandardized information that would not relate directly to the pronunciation of the word (such as brackets with hints on the meaning, multiple forms in a paradigm, etc.). Instead of trying to automatize this step, we worked through the data manually and also made extensive use of the possibility to correct lexical forms manually by keeping a list of lexemes that would later be replaced to a more standardized representations (available in the file etc/lexemes.tsv in the CLDF dataset).

3.4 CLDF Conversion

CLDF conversion was carried out in the "traditional" manner, following the Lexibank workflow of converting data to CLDF wordlists with CLDFBench (https://pypi.org/project/cldfbench, Forkel and List 2020). We profited from the flexibility of using custom Python code in CLDFBench by applying the concept filter at this stage, by checking for each word form, if its Concepticon concept set recurs in the Intercontinental Dictionary Series. Thanks to the Concepticon being available from CLDFBench, this amounts to one statement by which the dictionary is created, as shown in the code line from the Lexibank script below.

```
def cmd_makecldf(self, args):
    """ Convert the raw data to a CLDF dataset. """
    # select IDS concept list to check for concepts to be added
    ids = {c.concepticon_gloss for c in
            self.concepticon.conceptlists["Key-2016-1310"].concepts.values()
            if c.concepticon_gloss}
```

5 Conclusion

Although it seemed close to being impossible at first to provide a subset of the huge and impressive Dogon Comparative Wordlist, we found a way to address this task with standard procedures that fit nicely in the Lexibank workflow of converting datasets into CLDF Wordlists. In the future, we plan on linking the resulting concept list to Concepticon and carrying out an initial automatic analysis of the data, searching for cognate sets among the 25 Dogon varieties.

The data is curated on GitHub (https://github.com/languageislands/heathdogon) and archived on Zenodo (https://doi.org/10.5281/zenodo.8238983).

References

List, Johann-Mattis and Tjuka, Annika and van Zantwijk, Mathilda and Blum, Frederic and Barrientos Ugarte, Carlos and Rzymski, Christoph and Greenhill, Simon J. and Robert Forkel (2023): CLLD Concepticon [Dataset, Version 3.1.0]. Leipzig:Max Planck Institute for Evolutionary Anthropology.

Moran, Steven, Forkel, Robert, and Heath, Jeffrey (2016): Dogon and Bangime Linguistics. Jena:Max Planck Institute for the Science of Human History. https://dogonlanguages.org

Forkel, Robert and List, Johann-Mattis (2020): CLDFBench. Give your Cross-Linguistic data a lift. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation. 6997-7004.

Hantgan, Abbie and Babiker, Hiba and List, Johann-Mattis (2022): First steps towards the detection of contact layers in Bangime: A multi-disciplinary, computer-assisted approach [version 2; peer review: 2 approved]. Open Research Europe 2.10. 1-27.

Hantgan, Abbie (2022): The Small Bang. Computer-Assisted Language Comparison in Practice 5.12. 1-2. https://calc.hypotheses.org/5053

Hantgan, Abbie and List, Johann-Mattis (2022): Bangime: secret language, language isolate, or language island? A computer-assisted case study. Papers in Historical Phonology 7.1. 1-43.

Key, Mary Ritchie and Comrie, Bernard (2016): The Intercontinental Dictionary Series. Leipzig:Max Planck Institute for Evolutionary Anthropology. https://ids.clld.org

List, Johann-Mattis and Walworth, Mary and Greenhill, Simon J. and Tresoldi, Tiago and Forkel, Robert (2018): Sequence comparison in computational historical linguistics. Journal of Language Evolution 3.2. 130–144.

List, Johann-Mattis and Forkel, Robert and Greenhill, Simon J. and Rzymski, Christoph and Englisch, Johannes and Gray, Russell D. (2022): Lexibank, A public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9.316. 1-31. https://lexibank.clld.org

Moran, Steven and Cysouw, Michael (2018): The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles. Berlin:Language Science Press.

List, Johann-Mattis (2021): PySem: Python library for handling semantic data in linguistics. Leipzig:Max Planck Institute for Evolutionary Anthropology. https://pypi.org/projects/pysem

Sagart, Laurent and Jacques, Guillaume and Lai, Yunfan and Ryder, Robin and Thouzeau, Valentin and Greenhill, Simon J. and List, Johann-Mattis (2019): Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Science of the United States of America 116. 10317-10322.