Standardized Data Collections in Multilingual Computational Linguistics

Johann-Mattis List (University of Passau)

1 Background

Since we started the CLDF initiative in 2014, many datasets have been converted to CLDF, including word lists, structure datasets, and dictionaries. It is difficult to give a concrete number on the individual datasets that have been created, but it is quite likely that they exceed 200 or even 300 now. In order to increase the *findability* of CLDF datasets in the sense of the F in FAIR data (Wilkinson et al. 2016), we started to curate collections of individual CLDF datasets that we have released with Zenodo. The most prominent collection here is *Lexibank* (`https://zenodo.org/communities/lexibank`), offering standardized word lists. Another collection (so far without a Zenodo community) is the collection *CLDF Datasets* (`https://github.com/cldf-datasets`) which offers various kinds of data that are not primarily lexical, including phoneme inventories like Phoible (Moran and McCloy 2019) or the World Atlas of Language Structures Online (Dryer and Haspelmath 2013). A larger collection of digital dictionaries in CLDF is offered by Dictionaria (`https://dictionaria.clld.org`), and a larger collection of wordlists with numeral systems from the worlds' languages is offered by Numeralbank (`https://github.com/numeralbank/`). Additionally, certain types of legacy data which are often no longer expanded, have been given their own CLDF collection, including the still slightly growing Intercontinental Dictionary Series (`https://github.com/intercontinental-dictionary-series`, (Key and Comrie 2016)), or the datasets discussed in List (2014), which are now accessible in individual CLDF datasets (`https://github.com/sequenceComparison/`).

Why did it take such a long time to publish the first version of the Lexibank database?

2 Lexibank

Background

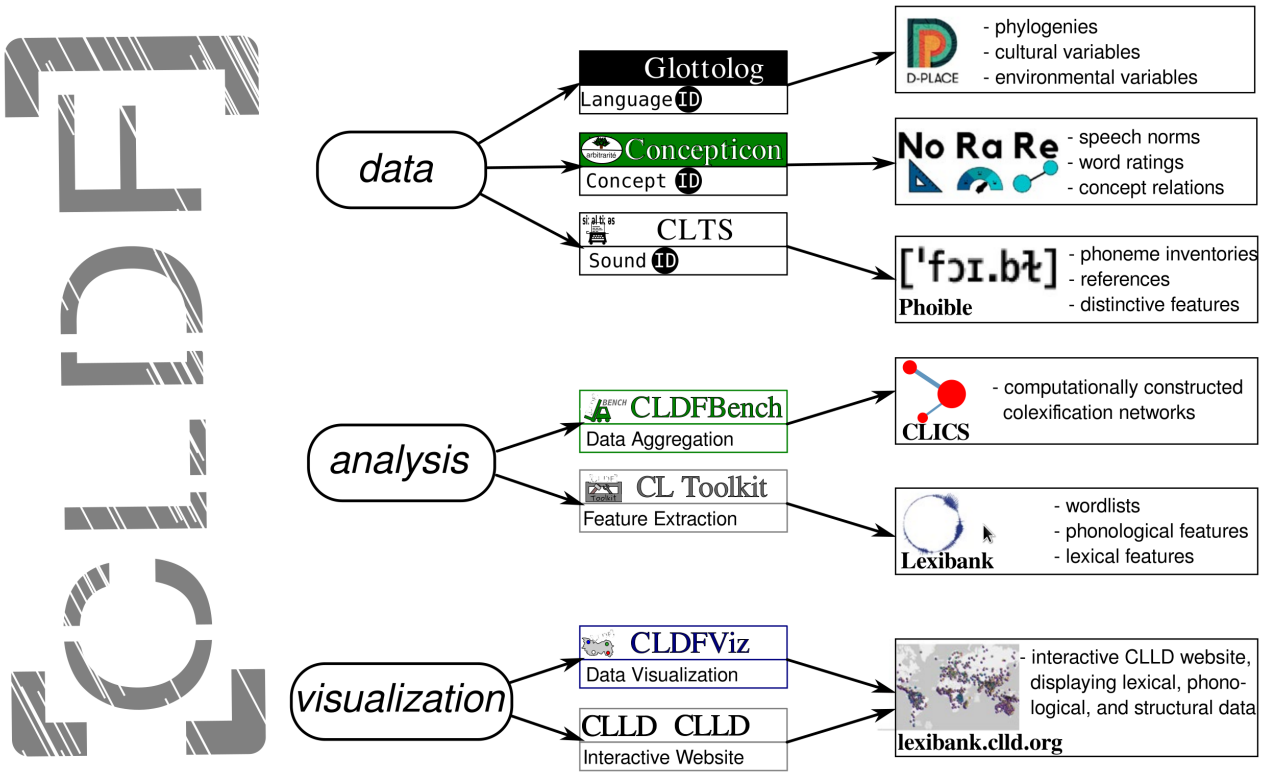
Lexibank is the largest collection of standardized wordlists in CLDF. Initiated in 2014, data collection reached its peak in 2018, after major components of the specific standards, like Concepticon (List et al. 2016) and CLTS (Anderson et al. 2018) had been published. Earliert test versions of Lexibank were published as part of the Database of Cross-Linguistic Colexifications (`https://clics.clld.org`), in 2018 (List et al. 2018) and in 2020 (Rzymski et al. 2020), which we will present in a separate session in more detail. In 2022, the first official version of Lexibank (Version 0.2) was published (List et al. 2022a). It consists of an aggregated CLDF dataset, in which data from 100 different CLDF datasets was aggregated and then consecutively analyzed, searching automatically for phonological and lexical features.

> Lexibank is a meta-collection of standardized wordlists compiled from various individual datasets. The standardized wordlists themselves are independently curated. Their curation fol- lows the data curation workflow of the Lexibank project, which uses the PyLexibank Python library (Forkel et al. 2021) to convert lexical data in custom formats into CLDF wordlists. The editorial board of the Lexibank project decides about the inclusion of individual datasets into the Lexibank wordlist collection. Datasets which are included in this collection need to be archived with Zenodo (`https://zenodo.org/`) and curated in a GIT repository (`https://github.com/`). Datasets included into the Lexibank wordlist collection are referenced with their Zenodo DOI and the URL of their GIT repository and classified for their level of standardization. (List et al. 2022a: 5/16)

What advantage has the automated search done in the context of Lexibank compared to a good traditional manual search in the grammatical literature?

Data Curation in Lexibank

The curation process of Lexibank data makes use of the CLDFBench package (Forkel and List 2020) that was extended by the PyLexibank plugin (Forkel et al. 2021). The major idea is to standardize the process by which an individual dataset is converted to CLDF as well as possible. This means that we start from the raw data, which may be manually adjusted, and then try to parse the data in order to read it into tables. Having done this, we convert the tabular data to CLDF, providing additional information on the concepts (which we manually or semi-automatically link to Concepticon), on the languages (which we manually link to Glottolog) and the phonetic transcriptions, which we segment and normalize at the same time by creating an orthography profile (Moran and Cysouw 2018) and applying it to the original transcriptions. The transcriptions can themselves be preprocessed with the help of code for the handling of lexical entries provided by PyLexibank.

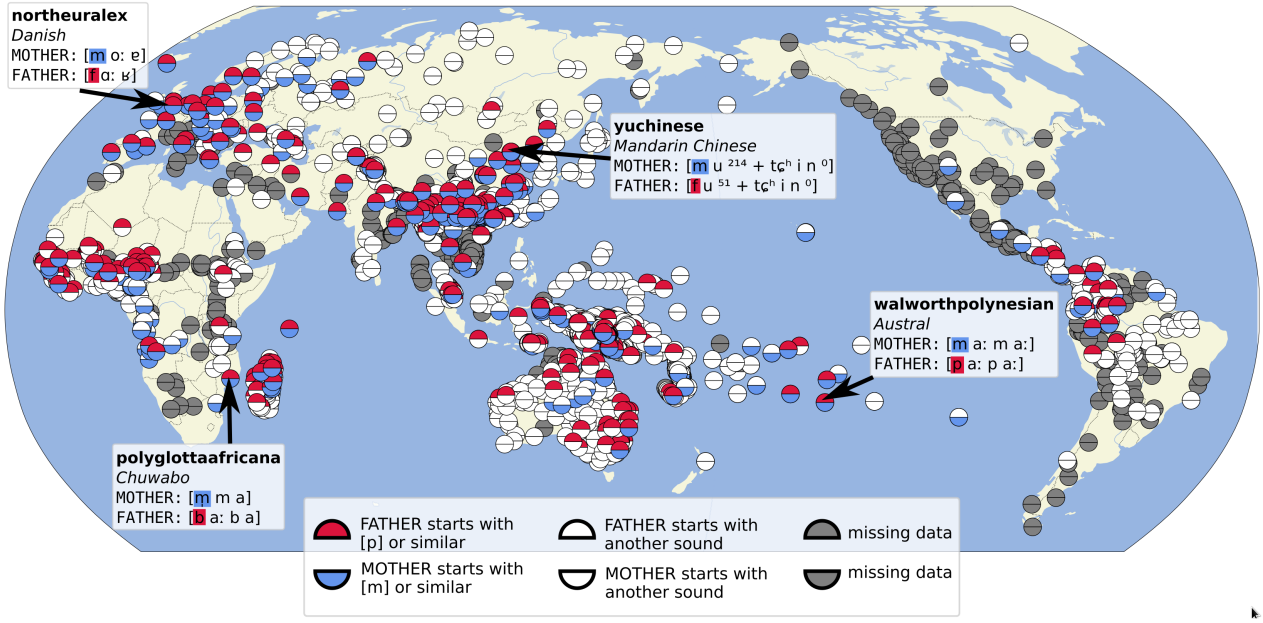


Reference catalogs have the advantage to allow us to access data from other databases that has been collected for different purposes, such as, for example, cultural data for language varieties or norm data for concepts. What research questions could one investigate with the extended access to cultural data and norm data from psychology?

Automated Feature Extraction

Once an aggregated word list is available, we can extract various phonological and lexical features automatically from the data. As an example, consider the feature observed by Jakobson (1960), who

discussed the observation made by many people and linguists, that many languages have a word form "mother" that starts with *m-* and often sounds like *mama* and a word for "father" that often starts with *p-* or *f-* and sounds like *papa*. Since our transcriptions are provided in the form of CLTS (Anderson et al. 2018, List et al. 2021a) and our concepts are linked to Concepticon (List et al. 2022b), we can easily formulate a query, in which we state that we search for words that start with an m-like sound for the concept MOTHER and with a p-like sound for the concept FATHER. The resulting data can then be easily plotted on a map with the help of the CLDFViz package (Forkel 2021).



What are the reasons that languages frequently choose *mama* and *papa* as words for "mother" and "father"?

Future of Lexibank

We are currently working on extended functionalities of Lexibank. Our main objective as of now is to look into fast database queries that can be applied to the extended Lexibank data (which will have many more phonetically transcribed language varieties in its upcoming version). The idea is that we directly convert the Lexibank CLDF data into a SQLITE database (or any other database system) and then query the data in order to answer specific questions. Using databases rather than CSV files has the advantage of speed, and as a result, many different queries can be "asked" quickly in order to test and generate hypotheses. Queries could even be asked on a website, given that the current version of all Lexibank data as an SQLITE database does not exceed 250 MB. Queries can output data in various forms. One could create a word list in LingPy's format (List et al. 2018) or the format required by EDICTOR (List 2017, List 2021). One can also create a CSV file with the values that would be sufficient to plot features of individual languages on a geographic map with the help of CLDFViz. All in all, we hope that queries in this form allow us to establish a Basic Linguistic Search Service (BLISS) that could play a similarly important role as the Basic Local Alignment Search Tool (BLAST) that revolutionized evolutionary biology (Altschul et al. 1990).

What kind of queries could we ask a Lexibank database?

3 CLDF Datasets

Background

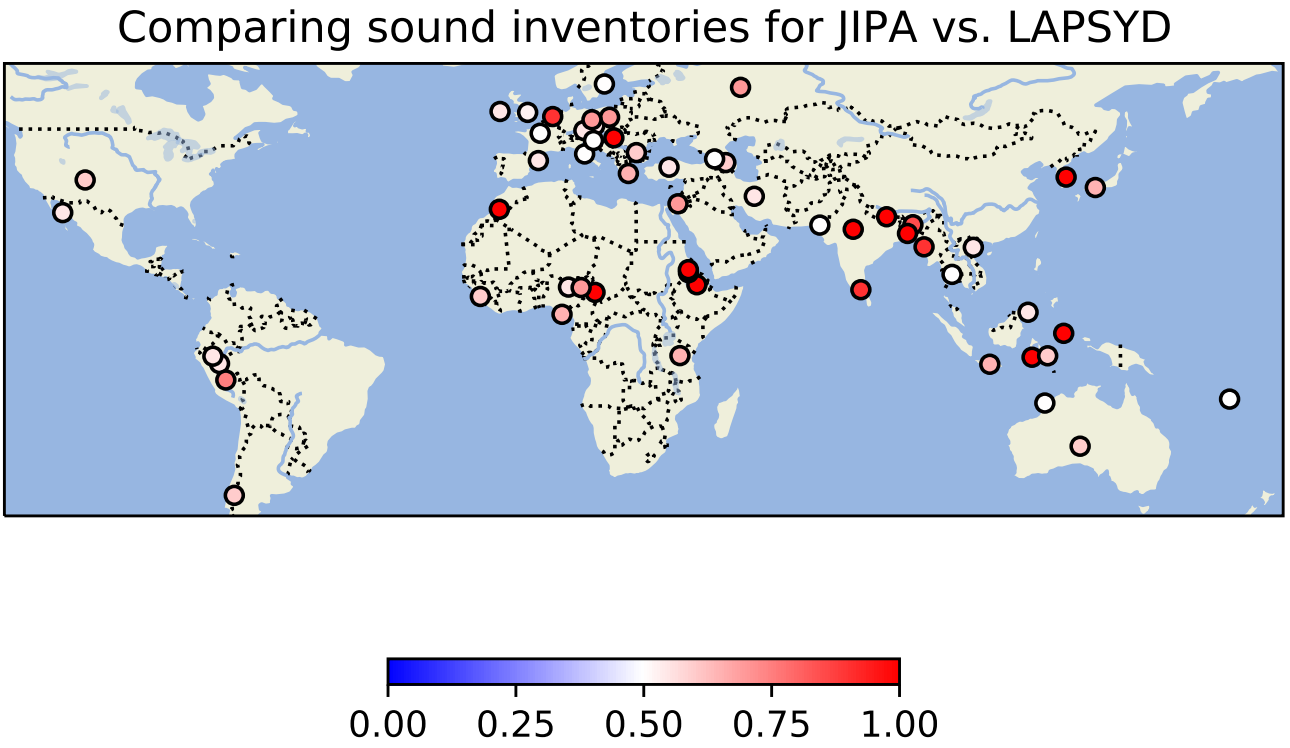
When starting to prepare lexical word lists in CLDF, we quickly realized that there are many other kinds of linguistic data that might also be worthwhile to be standardized. As a result, we began to prepare individual structural datasets in CLDF format, based on personal interests (List 2018). Later, it was decided to start collecting these datasets in a dedicated GitHub organization to not loose track of them. This organization, called CLDF Datasets (`https://github.com/cldf-datasets`) has still fewer repositories than the Lexibank organization, but it contains already 80 different public datasets as of today and is constantly growing. While adding datasets to the CLDF Datasets organization, we hope that we can identify specific sub types of data that might later also be either aggregated or compared.

What other kinds of cross-linguistic data apart from word lists could be similar enough to call for a consistent aggregation in a similar way in which this was done for Lexibank?

Phoneme Inventories

Based on our interest in different transcription systems, as reflected in the work on the Cross-Linguistic Transcription Systems reference catalog, we started to collect dedicated collections of phoneme inventories, trying to standardize them in a similar way in which we use the Concepticon to help to aggregate data from different word lists. While Concepticon offers identifiers for concepts, CLTS offers identifiers for speech sounds, and the idea was that we could use the identifiers in a similar way across datasets and in this way also point to differences in phonetic transcriptions across phoneme inventory collections.

Of the inventory datasets that are consistently linked to CLTS, we currently provide the LAPSyD database (Maddieson et al. 2013), the collection of phoneme inventories published in the Journal of the International Phonetic Association by Baird et al. (2021), and the Eurasian inventories by Nikolaev et al. (2017). Databases like Phoible (Moran and McCloy 2019) are available in CLDF and referenced by CLTS but do not yet include the explicit link to a given version of the CLTS reference catalog themselves. We estimate that there is a potential to link at least 10 more datasets in a similar form to CLTS. Thus, the currently still small collection could further grow in the future and allow for additional possibilities to build on initial studies that compare how well phoneme inventories for the same varieties correspond when collected by different authors (Anderson et al. 2021).



The map above (from Anderson et al. 2021) compares phoneme inventory sizes in the JIPA collection by Baird et al. (2021) and the LAPSyD database by Maddieson et al. (2013). Is there any pattern that can be detected with respect to the differences in the phoneme inventory sizes?

Structure Datasets

Structure datasets are very diverse by their nature, and we find collections of very specific features, such as "The third person pronoun is tā, or cognate to it." for Chinese dialect varieties (Norman 2003, `https://github.com/cldf-datasets/normansinitic`) or "Does the language have morphosyntactic plural markers?" (Tang and Her 2019, `https://github.com/cldf-datasets/tangclassifiers`). As a result, comparing individual datasets that have been collected so far is much more difficult if not impossible, than comparing word lists or phoneme inventory collections. In order to approach the problem, a metadata catalogue of structural properties of languages would be needed, and this catalogue would have to identify those features which frequently recur across the languages in the world.

While such an enterprise has not been undertaken yet, our work on the Lexibank project has initiated a first step into this direction. In automatically computing lexical and phonological features for the data in the Lexibank collection, we take direct inspiration from the features in the World Atlas of Language Structures Online (WALS, Dryer and Haspelmath 2013), and our feature computation workflow also notes similarities among the features we compute and their counterparts in the WALS database. Our idea was to further expand these collections of automatically computed features and to note more clearly and systematically which databases provide features that have been manually collected. In this way, a future reference catalog, albeit a small one to begin with, could well be prepared in the nearer future and also provide concrete information on the computability or the computability status of certain features that have been collected for the world's languages.

No.	Identifier	Name	Type
1	LegAndFoot	has the same word form for foot and leg	colexification
2	ArmAndHand	arm and hand distinguished or not	
3	BarkAndSkin	bark and skin distinguished or not	
4	FingerAndHand	finger and hand distinguished or not	
5	GreenAndBlue	green and blue colexified or not	
6	RedAndYellow	red and yellow colexified or not	
7	ToeAndFoot	toe and foot colexified or not	
8	SeeAndKnow	see and know colexified or not	
9	SeeAndUnderstand	see and understand colexified or not	
10	ElbowAndKnee	elbow and knee colexified or not	
11	FearAndSurprise	fear and surprise colexified or not	
12	CommonSubstringInElbowAndKnee	elbow and knee are partially colexified or not	partial colexification
13	CommonSubstringInManAndWoman	man and woman are partially colexified or not	
14	CommonSubstringInFearAndSurprise	fear and surprise are partially colexified or not	
15	CommonSubstringInBoyAndGirl	boy and girl are partially colexified or not	
16	EyeInTear	eye partially colexified in tear	affix colexification
17	BowInElbow	bow partially colexified in elbow	
18	CornerInElbow	corner partially colexified in elbow	
19	WaterInTear	water partially colexified in tear	
20	TreeInBark	tree partially colexified in bark	
21	SkinInBark	skin partially colexified in bark	
22	MouthInLip	mouth partially colexified in lip	
23	SkinInLip	skin partially colexified in lip	
24	HandInFinger	hand partially colexified in finger	
25	FootInToe	foot partially colexified in toe	
26	ThreeInEight	three partially colexified in eight	
27	ThreeInThirteen	three partially colexified in thirteen	
28	FingerAndToe	finger and toe colexified or not	
29	HairAndFeather	hair and feather colexified or not	
30	HearAndSmell	hear and smell colexified or not	

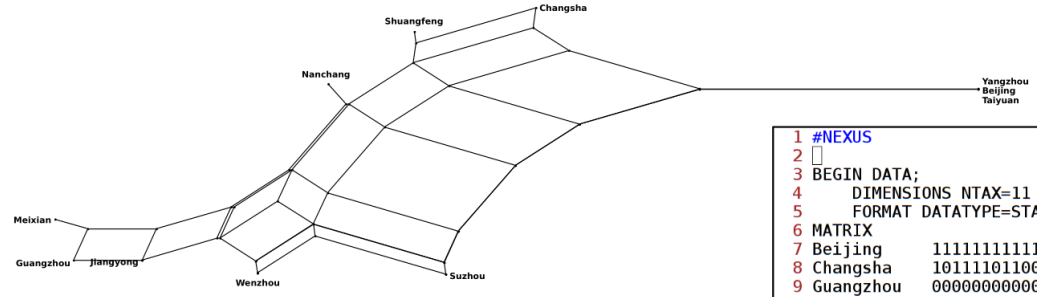
The table above shows lexical features computed from Lexibank data. What is meant with *partial colexification* and with *affix colexification*?

4 Combined Data

As a final example for our current efforts in working with CLDF and testing the limits of the format specification, we have started to create combined datasets in which we combine, e.g., features with words in a wordlist. Thus, we often find lexical word lists accompanied by phoneme inventories in the literature. When digitized, we can render both datasets in one combined CLDF datasets in which the language table is shared among both datasets.

As of now, we experiment with the combination of structure datasets with word lists, the combination of dictionaries with word lists (where a word list can be derived from a dictionary), and with the modeling of more complex features, such as, for example, colexifications (which we will discuss in an upcoming

session). As a most complete way of combining data for a single resource, we hope to integrate interlinear-glossed text with additional resources, such as, for example, phoneme inventories, word lists, and even dictionaries in one single CLDF package. So far, however, we have not found data collections which would offer all these resources in combination (an initial example is discussed in List et al. 2021b). We will discuss how texts and corpus data can be handled in CLDF in an additional session.



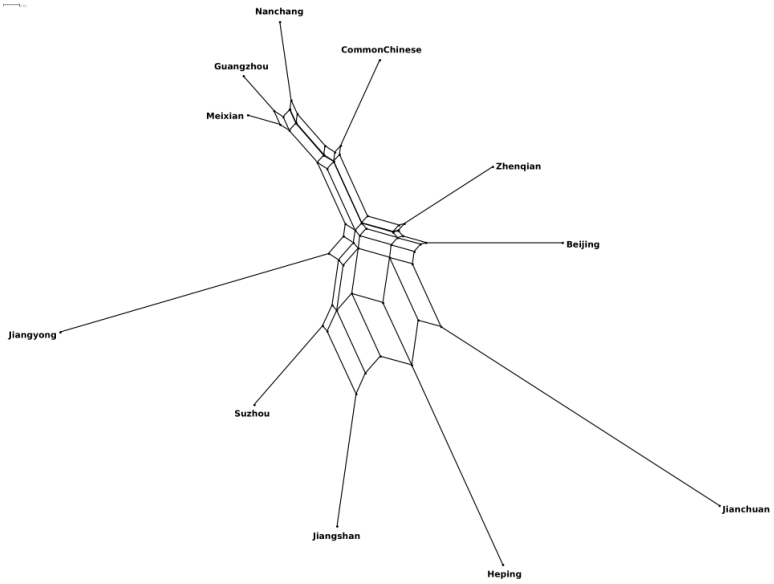
(A) Neighbor-Net drawn from structural data

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=11 NCHAR=15;
    FORMAT DATATYPE=STANDARD GAP=- MISSING=?;
MATRIX
Beijing     111111111111111
Changsha    101111011000010
Guangzhou   000000000000000
Jiangyong   000100000000000
Meixian     000000100000000
Nanchang    001101001000010
Shuangfeng  101101011000000
Suzhou      001101000110001
Taiyuan     111111111111111
Wenzhou     001100000010000
Yangzhou    111111111111111

;
END;
```

(C) NEXUS format for structural data



(B) Neighbor-Net drawn from lexical data

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=11 NCHAR=102;
    FORMAT DATATYPE=STANDARD GAP=- MISSING=?;
MATRIX
Beijing        10000100001101010101001011001
CommonChinese  01000010001011010101001010110000
Guangzhou      01000010001101010101000110101000
Heping         10000001001101010100110010110000
Jianchuan      00100000101100110101010001100010
Jiangshan      00010100001011010101010010110000
Jiangyong      00001000011101010011000110101000
Meixian        01000010001101010101000110101000
Nanchang       01000010001101001101000110110000
Suzhou         00010100001101010101001010110000
Zhenqian       10000000101101010101001010110000
;
END;
```

(D) NEXUS format for lexical data (excerpt)

Why are the Neighbor-net representations of the data by Norman (2003), as described in Forkel and List (2020) so different?

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). "Basic local alignment search tool." *Journal of Molecular Biology* 215.3, 403 –410.

Anderson, C., T. Tresoldi, T. C. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznań Linguistic Meeting* 4.1, 21–53.

Anderson, C., T. Tresolid, S. J. Greenhill, R. Forkel, R. Gray, and J.-M. List (2021). *Measuring variation in phoneme inventories*.

Baird, L., N. Evans, and S. J. Greenhill (2021). "Blowing in the wind: Using 'North Wind and the Sun' texts to sample phoneme inventories." *Journal of the International Phonetic Association* 0.0, 1–42.

Dryer, M. S. and M. Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Forkel, R. (2021). "CLDFViz. A python library providing tools to visualize data from CLDF datasets [Software Library, Version 0.5.0]."

Forkel, R., S. J. Greenhill, H.-J. Bibiko, C. Rzymski, T. Tresoldi, and J.-M. List (2021). *PyLexibank. The python curation library for lexibank [Software Library, Version 2.8.2]*. Geneva: Zenodo.

Forkel, R. and J.-M. List (2020). "CLDFBench. Give your Cross-Linguistic data a lift." In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*. "LREC 2020" (Marseille). Luxembourg: European Language Resources Association (ELRA), 6997–7004.

Jakobson, R. (1960). "Why 'Mama´and 'Papa?." In: *Perspectives in psychological theory: Essays in honor of Heinz Werner*. Ed. by B. Kaplan and S. Wapner. New York: International Universities Press, 124–134.

Key, M. R. and B. Comrie (2016). *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://ids.clld.org.
List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
— (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, 9–12.
— (09/03/2018). "Representing Structural Data in CLDF." *Computer-Assisted Language Comparison in Practice* 08.08.
— (2021). *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets. Version 2.0.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://digling.org/edictor.
List, J.-M., C. Anderson, T. Tresoldi, and R. Forkel (2021a). *Cross-Linguistic Transcription Systems. Version 2.1.0*. Jena: Max Planck Institute for the Science of Human History. URL: https://clts.clld.org.
List, J.-M., M. Cysouw, and R. Forkel (2016). "Concepticon. A resource for the linking of concept lists." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. "LREC 2016" (Portorož, 05/23–05/28/2016). Ed. by N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis. Luxembourg: European Language Resources Association (ELRA), 2393–2400.
List, J.-M., R. Forkel, S. J. Greenhill, C. Rzymski, J. Englisch, and R. D. Gray (2022a). "Lexibank, A public repository of standardized wordlists with computed phonological and lexical features." *Scientific Data* 9.316, 1–31.
List, J.-M., N. A. Sims, and R. Forkel (2021b). "Towards a sustainable handling of interlinear-glossed text in language documentation." *ACM Transactions on Asian and Low-Resource Language Information Processing* 20.2, 1–15.
List, J.-M., A. Tjuka, C. Rzymski, S. J. Greenhill, and R. Forkel (2022b). *CLLD Concepticon [Dataset, Version 3.0.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://concepticon.clld.org/.
List, J.-M., M. Walworth, S. J. Greenhill, T. Tresoldi, and R. Forkel (2018). "Sequence comparison in computational historical linguistics." *Journal of Language Evolution* 3.2, 130–144.
Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino. (2013). "LAPSyD: Lyon-Albuquerque Phonological Systems Database." In: *Proceedings of Interspeech*. (Lyon, 08/25–08/29/2013).
Moran, S. and M. Cysouw (2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Berlin: Language Science Press.
Moran, S. and D. McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.
Nikolaev, D., A. Nikulin, and A. Kukhto (2017). *The database of Eurasian phonological inventories*. Moscow: RGGU.
Norman, J. (2003). "The Chinese dialects. Phonology." In: *The Sino-Tibetan languages*. Ed. by G. Thurgood and R. J. LaPolla. London and New York: Routledge, 72–83.
Rzymski, C. et al. (2020). "The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies." *Scientific Data* 7.13, 1–12. URL: https://clics.clld.org.
Tang, M. and O.-S. Her (2019). "Insights on the Greenberg-Sanches-Slobin generalization: Quantitative typological data on classifiers and plural markers." *Folia Linguistica* 53.2, 297–331.
Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3, 1–9.