Cross-Linguistic Data Formats

Johann-Mattis List (University of Passau)

1 Introduction

Data in Linguistics

Linguistics is a discipline in which data play an important role. The main part of the work of many linguists consists in the inspection of data, in the curation of data, in the analysis of data, or in the correction of data. We need grammatical data to investigate grammatical phenomena. These include, among others, example sentences from larger corpora, usually presented in the form of interlinear-glossed text (Lehmann 2004). We need typological data in order to investigate questions on the structure of the languages in the world. These data are typically larger collections of phenomena extracted from individual grammars. If we want to investigate the lexicon of languages, we need wordlists or dictionaries. If data are not available, one needs to create one's own datasets, for example, by going to the field and searching for informants of a given language variety, or by inspecting secondary sources from which data could be extracted.

Are there any fields of linguistics in which data do not play a role?

Data in Comparative Linguistics

People working in the field of language comparison are traditionally even hungrier for data than people working on one particular language's syntax. When comparing languages, we cannot create the data in our heads through introspection. In order to investigate phenomena like language change, we need to compare different data points on the same or different language varieties, and these data points cannot be generated in our heads, they need to be collected. The process of data collection in the field of comparative linguistics may turn out to be quite tedious. Comparative linguists – specifically those working in traditional paradigms – sift through dictionaries, word lists, historical documents, grammars, they interview informants in order to gain more and fresh data on particular language varieties that are not very well documented, and they normally spend a much larger time of their research on the collection of data than on anything else. The results of studies on comparative linguistics can be shared in multiple forms. Etymological dictionaries, for example, are considered to be the "king's discipline" in historical linguistics, because they allow us to see the development of one particular language or an entire language family. In linguistic typology, the major research output are books devoted to specific specific topics of grammar that can then be investigated in the form of a survey, such as, for example, "number" (Corbett 2004), have for a long time been the major research output. Nowadays, with the advent of larger online collections that can be searched on the internet, another major research output are typological databases, which are typically collected by individuals reading the grammars for particular languages in order to extract certain aspects of information. A famous example for this kind of data is the *World Atlas of Language Structures Online* (Dryer and Haspelmath 2013). The problem of etymological dictionaries is that they are still delivered in the form of a book. Although knowledge has been collected in a systematic manner in order to compile them, the knowledge is no longer available in a systematic form, once the dictionary has been compiled. On the contrary, in order to work with etymological dictionaries, the only way to use them in many cases is to inspect them manually, reading individual entries and digesting their content. While typological databases allow us to search quickly for one specific phenomenon, they often go too far in the way in which the original data has been converted to fit the format of the target database. As a result, it is not always useful to rely on the information blindly, and those who have been working with these databases know very well, that there is often no way around reading the original literature from which these collections have been compiled.

Etymological dictionaries are often based on older literature, which is frequently quoted, remixed, and modified. Where do we also find this attempt to cumulatively bring the knowledge about some topic to perfection?

Data Problems in Comparative Linguistics

There are numerous problems resulting from the way in which data is managed and organized in the field of comparative linguistics. We can distinguish three major problems. The problem of (a) availability, (b) transparency, and (c) comparability.

The lack of availability is very annoying, not in the sense that we have no access to a given article in the form of a scan or a book, but rather because many authors collect data, write articles about them, but then do not share their data officially. It is still not surprising that articles are being published in which new ideas are postulated or new conclusions are being made, but in which scholars do not share the data upon which they base their conclusions openly (Tamburelli and Brasca 2017). The same holds for many grammatical descriptions, in which scholars extract individual sentences from their personally collected private corpus but never reference them sufficiently, nor offer the full corpus. This can be seen from the following quote taken from a review of a handbook on Sino-Tibetan languages.

> It is disappointing that so many among the authors of newly commissioned articles did not cite their data; this failing is particuarly perplexing in the case of those authors who benefited from the generosity of agencies that explicitly require archiving in public repositories. The move toward open data is still in its early days. (Hill 2017: 306)

Apart from the availability we also face the problem of data *transparency*. As an example, see Bengtson (2017), where the author tries to show the readers that Basque and North Caucasian are related.

(gloss)	Basque	Chechen	Avar	Lak / Dargi	Lezgi	Prot-West-Caucasian	Proto-North-Caucasian
die	*hil	=al-	=al'=	L =ič'a D -ibk'-	q'i-	* ƛ̣ə - / *ƛ̣a-	* =iwƛ̣E
dog	*hor	pħu 'male dog'	hoy	D χa	χor (Budukh)	*ŁIwa	*χHwĕy-rV-
ear	*be=laři	ler-g		D liħi		*ŁA-	*łĕHi
fire	*śu	ts'e	ts'a	L ts'u D ts'a	ts'ay	*mA=c w̦ a	*c̣ ăyɨ
horn	*a=dař	kur	tł:ar		firi 'mane'		PEC *ƛwɨ̆ rV
I	*ni			L na D nu		*q̇:IwA 'to hear; to be heard'	* =ĭq̇Ē

Here, it is incredibly hard to interpret or understand the similarities which the author claims to have detected.

As a last problem, we have the problem of comparability of research data. Here, we often find the situation that scholars do not pay attention to sharing their data in such a form that they could be easily compared with other, often similar data, published in independent studies. It is clear that comparability of data is hard to achieve, but some basic aspects of comparability, like a consistent indication of the origin of data, a unified phonetic transcription, consistent standards in naming language varieties and concepts, all of this is indispensable if we want to contribute with our data to science in general. Comparability is unfortunately mostly ignored in comparative linguistics, although many scholars appreciate large data collections in which data have been made comparable. The lack of comparability also contributes to the increasing problem that studies in comparative linguistics can often not be reproduced.

In which cases would it be justified or even important *not* to share research data?

2 The CLDF Initiative

General Ideas

The *Cross-Linguistic Data Formats*-Initiative (CLDF, Forkel et al. 2018, `http://cldf.clld.org`) has the following goals:

(a) working toward the standardization (and retro-standardization) of cross-linguistic research data,

(b) establishing software APIs that help us to check if data conform to these standards and to make use of the data in one's research, and

(c) providing examples for *best-practice*.

In order to address (a), CLDF proposes to make use of metadata bases (*reference catalogs*) like Glottolog (Hammarström et al. 2021), Concepticon (List et al. 2022b), and CLTS (List et al. 2021a). These metadata collections help scholars to make explicit what kind of data they use (which language varieties, which concepts, which sounds). Their goal is to contribute to increasing the *comparability of research data* in comparative linguistics.

In order to address (b), CLDF provides software packages (typically written in Python) that can be used to access data coded in CLDF (CL Toolkit, `https://pypi.org/project/cltoolkit`, List and Forkel 2021), to convert existing data to CLDF (CLDFBench, `https://pypi.org/project/cldfbench`, Forkel and List 2020), or to check if a given dataset conforms to the standards outlined by CLDF (PyCLDF, `https://pypi.org/project/pycldf`, Forkel et al. 2021b). The software in this contexts makes sure that data are both machine- and human-readable at the same time.

In ordert to accomplish (c), CLDF propogates collections of existing datasets coded in CLDF. These collections can be used and inspected by users interested to present their own data in CLDF. They give concrete examples of problem-handling within the CLDF framework and serve as a practial knowledge base where users can take inspiration for their own work. The by now largest collection of individual CLDF datasets, all prepared with the help of the CLDFBench package is the Lexibank repository, offering more than 100 datasets consisting of CLDF wordlists, covering several thousand of the worlds' languages and several dozens of the world's language families (List et al. 2022a).

What is the advantage of using metadata collections like Glottolog when collecting data transparently?

Technical Aspects

The technical aspects of CLDF can be retrieved from the project website (`http://cldf.clld.org`), where one finds a specification and individual examples of the underlying ontology. Currently, CLDF offers three major datatypes, namely Wordlist, Structure Dataset, and Dictionary. The general format in which tabular data are shared is CSV (comma-separated value) with an additional metadata file in JSON format that explains how the CSV data should be interpreted and which columns are linked with each other, following the W3C recommendations for tabular data and metadata on the web (W3C

Consortium 2015, `https://csvw.org`). The CLDF ontology builds on the *General Ontology for Linguistic Description* (GOLD, Community 2010). The `pycldf` Python package (`https://github.com/glottobank/pycldf`, Forkel et al. 2021b) provides the possibility to read and write CLDF data, and also includes commandline facilities to check of a dataset conforms to the CLDF requirements as well as to convert a CLDF dataset into SQLITE format (a very common format for databases that can be read from normal files). The `CLDFBench` package (Forkel and List 2020), allows to convert data to CLDF in a convenient way, using the commandline and standardized Python code. CLDFBench has been extended with `PyLexibank` (Forkel et al. 2021a), a Python package dedicated to the creation of CLDF Wordlists used for the creation of the Lexibank repository (List et al. 2022a).

Why use tabular formats if you could use TEI or plain XML?

Standards in CLDF

CLDF consists of different modules in which specific standard requirements for certain data types are stored. As of now, there are three main modules (a) Wordlist, (b) Dictionary, and (c) Structure Dataset. Additional examples exist that show how more complex data types can also be represented in CLDF, including interlinear-glossed text (List et al. 2021b), and combined datasets in which a wordlist is accompanied by a structure dataset or in which particular structural datasets, like phoneme inventories are handled in a similar form, which could later on be modeled in their own module (Anderson et al. 2021).

In order to convert one's data to CLDF, the first step is to select the appropriate data model (the module). If no model fits a given requirement, one can also use a Generic module that has minimal basic requirements. Most linguistic data come along in the form of *triples*, consisting of a language (variety), an parameter (the question that a dataset asks), and a value (the answer regarding the question). Thus, if one creates a dataset that asks whether a language has an article or not, one would start from a list of individual language varieties, then ask the question (the parameter) "has article?", and then provide the answer "yes / no / dunno". This triplet structure could in theory be rendered by a simple table, rendering this triple structure.

Language_ID	Parameter_ID	Value
German	has article?	yes
English	has article?	yes
Chinese	has article?	no

However, since we may want to provide additional information on the languages in our sample, we'd prefer to add an individual table for the languages, where this information is stored. Additionally, we may want to add more information on the parameter (or the collection of multiple parameters), and this information would then also better be stored in a specific parameter table. Finally, if one wants to store the sources (e.g., the grammar from which one has taken the information on the article status) one would want to provide them as well in a separate file.

As a result, a typical Structure Dataset in CLDF can consist of a language table, a parameter table, and a value table, and a list of sources (in BibTeX format) which are linked with each other via identifiers.

The same model can be used – with slight modifications – to account for a word list, where we have again one table for the languages, one table for the parameters (the concepts in this specific case) and one table for the values (the word forms, called form table in CLDF).

Language_ID	Parameter_ID	Form
German	HAND	hant
English	HAND	hænd
Chinese	HAND	ʃɔu^{214}
German	FOOT	fuːs
English	FOOT	hænd
Chinese	FOOT	tsu^{35}

What is the difference between a word list and a dictionary?

References

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

Bengtson, J. D. (2017). *The Euskaro-Caucasian Hypothesis. Current model. A proposed genetic relationship between Basque (Vasconic) and the North Caucasian language family*. Ed. by A. for the Study of Language in Prehistory.

Community, G. (2010). *General Ontology for Linguistic Description (GOLD)*. Ontology. Department of Linguistics (The LINGUIST List), Indiana University.

Corbett, G. G. (2004). *Number*. Cambridge: Cambridge University Press.

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

Forkel, R., S. J. Greenhill, H.-J. Bibiko, C. Rzymski, T. Tresoldi, and J.-M. List (2021a). *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.

Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). “Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics.” *Scientific Data* 5.180205, 1–10.

Forkel, R., C. Rzymski, and S. Bank (2021b). *PyCLDF (Version 1.18.0)*. Jena: Max Planck Institute for the Science of Human History.

Hammarström, H., M. Haspelmath, R. Forkel, and S. Bank (2021). *Glottolog. Version 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://glottolog.org`.

Hill, N. W. (2017). “The State of Sino-Tibetan.” Review of Thurgood and Lapolla (2017) The Sino-Tibetan Languages. Second Edition. *Archiv Orientální* 85, 305–315.

Lehmann, C. (2004). “Interlinear morphemic glossing.” In: *Morphology. An international handbook*. Ed. by G. E. Booij, C. Lehmann, J. Mugdan, and S. Skopeteas. Vol. 2. Berlin and New York: De Gruyter, 1834–1857.

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. and R. Forkel (2021). *CL Toolkit. A Python Library for the Processing of Cross-Linguistic Data [Software Library, Version 0.1.1]*. Geneva: Zenodo.

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/`.

Tamburelli, M. and L. Brasca (2017). “Revisiting the classification of Gallo-Italic: a dialectometric approach.” *Digital Scholarship in the Humanities* fqx41.

W3C Consortium (12/17/2015). *Model for Tabular Data and Metadata on the Web*. W3C Recommendation. W3C.