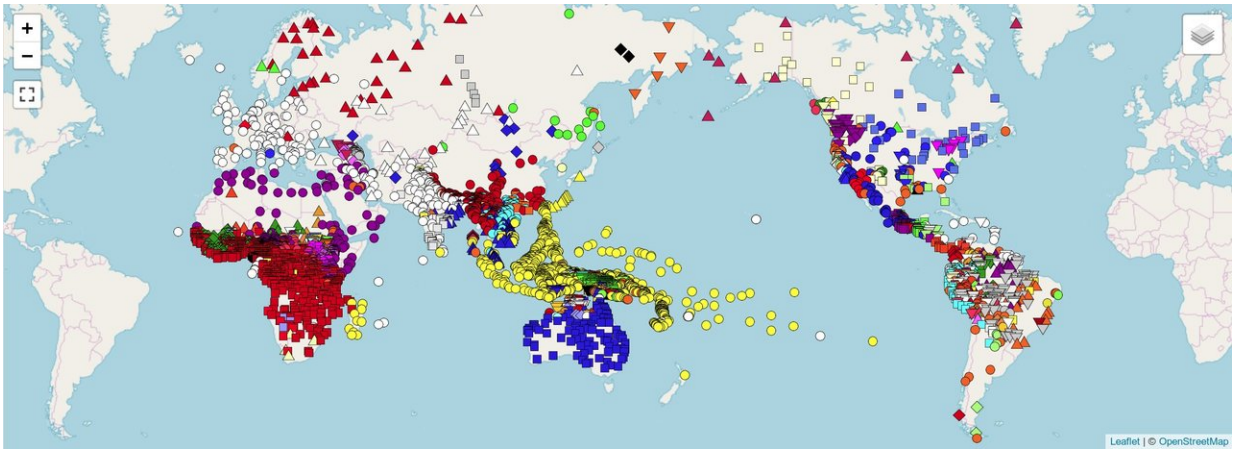
Guidelines for a standardized data format for use in cross-linguistic studies

October 16 2018



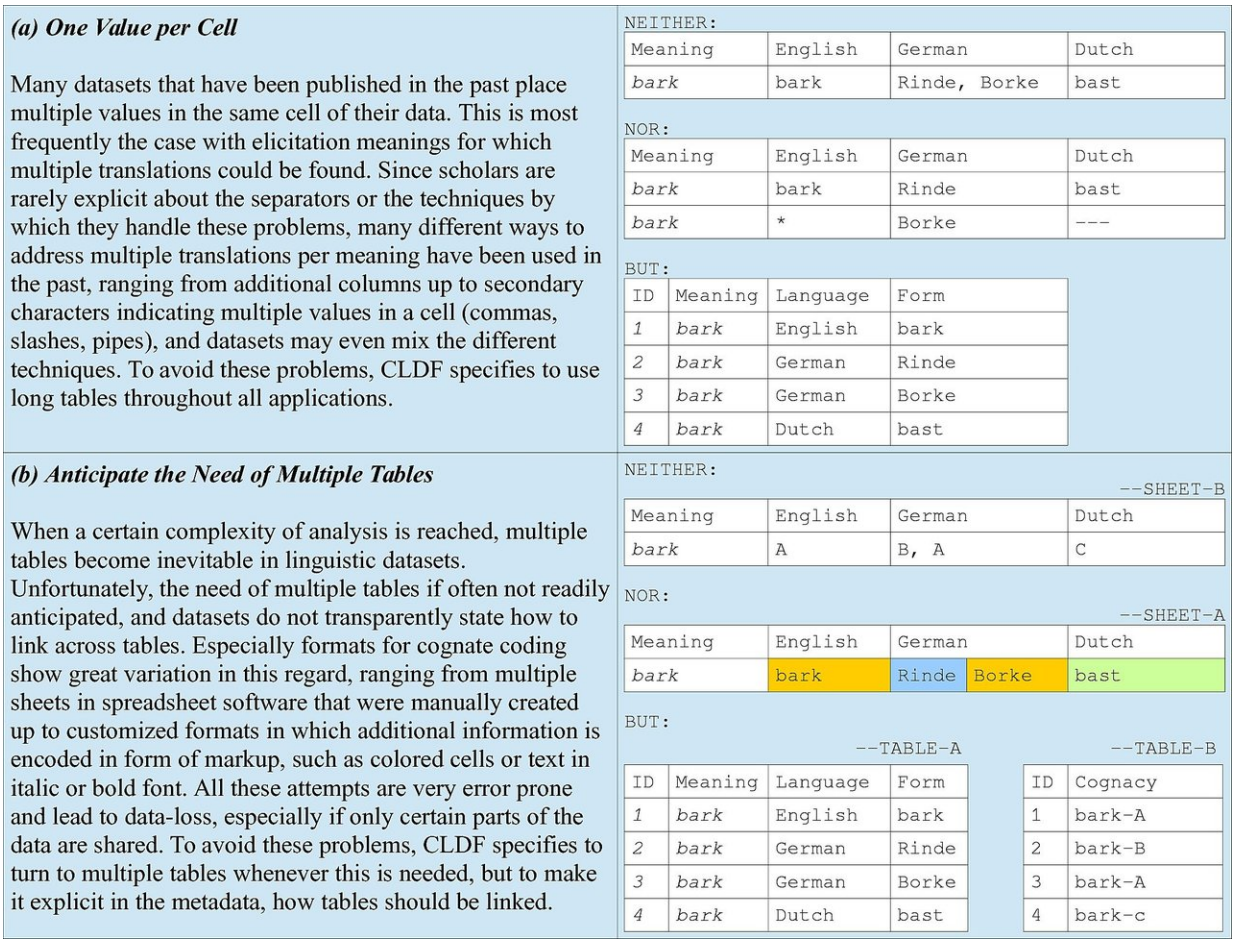
A world map showing data points, for which the researchers plan to gather unified data (e.g., data that is directly comparable) using the guidelines given in the paper. Credit: OpenStreetMap. Forkel et al. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*.

An international team of researchers, members of the Cross-Linguistic Data Formats Initiative (CLDF) led by the Max Planck Institute for the Science of Human History, has proposed new guidelines on cross-linguistic data formats in order to facilitate sharing and data comparisons between the growing number of large linguistic databases worldwide. This format provides a software package, a basic ontology and usage examples.

There is an increasing number of linguistic databases worldwide, raising the possibility of a vast network for potential comparative studies. However, these databases are generally created independently of each other, and often have a unique and narrow focus. This means that the formats used for encoding the data are often different, creating difficulties in comparing data across databases.

The Cross-Linguistic Data Formats Initiative (CLDF) is an effort to resolve these issues. In a paper published in *Scientific Data*, the CLDF sets out proposed guidelines for a standardized format for linguistic databases, and also supplies a software package, a basic ontology and usage examples of best practices. The goal of this effort is to facilitate sharing and re-use of data in comparative linguistics.

The CLDF provides a data model underlying its recommendations that aims to be simple, yet expressive, and is based on the data model previously developed for the Cross-Linguistic Data project. This model has four main entities: (a) languages; (b) parameters; (c) values; and (d) sources. In the model, each value is related to a parameter and a language, and can be based on multiple sources. There are additionally references for sources, and references can also have contexts (which, for example, for printed references would be page numbers).



(a) One Value per Cell

Many datasets that have been published in the past place multiple values in the same cell of their data. This is most frequently the case with elicitation meanings for which multiple translations could be found. Since scholars are rarely explicit about the separators or the techniques by which they handle these problems, many different ways to address multiple translations per meaning have been used in the past, ranging from additional columns up to secondary characters indicating multiple values in a cell (commas, slashes, pipes), and datasets may even mix the different techniques. To avoid these problems, CLDF specifies to use long tables throughout all applications.

NEITHER:

Meaning	English	German	Dutch
bark	bark	Rinde, Borke	bast

NOR:

Meaning	English	German	Dutch
bark	bark	Rinde	bast
bark	*	Borke	---

BUT:

ID	Meaning	Language	Form
1	*bark*	English	bark
2	*bark*	German	Rinde
3	*bark*	German	Borke
4	*bark*	Dutch	bast

(b) Anticipate the Need of Multiple Tables

When a certain complexity of analysis is reached, multiple tables become inevitable in linguistic datasets. Unfortunately, the need of multiple tables if often not readily anticipated, and datasets do not transparently state how to link across tables. Especially formats for cognate coding show great variation in this regard, ranging from multiple sheets in spreadsheet software that were manually created up to customized formats in which additional information is encoded in form of markup, such as colored cells or text in italic or bold font. All these attempts are very error prone and lead to data-loss, especially if only certain parts of the data are shared. To avoid these problems, CLDF specifies to turn to multiple tables whenever this is needed, but to make it explicit in the metadata, how tables should be linked.

NEITHER:

--SHEET-B

Meaning	English	German	Dutch
bark	A	B, A	C

NOR:

--SHEET-A

Meaning	English	German		Dutch
bark	bark	Rinde	Borke	bast

BUT:

--TABLE-A

ID	Meaning	Language	Form
1	*bark*	English	bark
2	*bark*	German	Rinde
3	*bark*	German	Borke
4	*bark*	Dutch	bast

--TABLE-B

ID	Cognacy
1	bark-A
2	bark-B
3	bark-A
4	bark-c

Basic rules of data coding included in the guidelines, taking cognate coding in wordlists as an example. (a) illustrates why long tables should be favored throughout all applications. (b) underlines the importance of anticipating multiple tables along with metadata indicating how they should be linked. Credit: Forkel et al. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*.

The CLDF data model is a package format in which a dataset would be made up of a set of data files containing tables, and a descriptive file that defines the relationships between the tables. Each linguistic data type would have a CLDF module and additional components, which would be the aspects of the data in the module that recur across multiple data

types. The CLDF modules would also contain terms from the CLDF ontology. The ontology is a list of vocabulary that represents objects and properties with well-known semantics in comparative linguistics. This makes it possible for users to reference these terms in a uniform way.

A software package to enable validation and manipulation

The CLDF specifications use common file formats—such as CSV, JSON and BibTeX—that are widely supported, with the goal that these files can easily be read and written on many platforms. Even more importantly, the standardized format will allow researchers without programming skills to access and manipulate the data with preexisting tools, to avoid restricting the package only to researchers with sufficient programming skills to create their own tools. To facilitate this, the CLDF has created a "cookbook" repository for scripts for use with the CLDF specifications.

"We want to bring access to these data and the ability to compare them to as many researchers as possible," says Johann-Mattis List of the Max Planck Institute for the Science of Human History. Robert Forkel, one of the driving forces behind the CLDF initiative, also notes that the CLDF format is not limited to linguistic data alone, but can also incorporate databases of cultural and geographic data, for example. "CLDF may drastically facilitate the testing of questions regarding the interaction between linguistic, cultural, and environmental factors in linguistic and cultural evolution."

More information: Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics , *Scientific Data*, DOI: 10.1038/sdata.2018.205

Provided by Max Planck Society