Reference Catalogs

Johann-Mattis List (University of Passau)

1 Background

Reference catalogs – in the sense in which we use them in CLDF – are understood as bigger metadata collections that are curated by a team of dedicated scholars independently of individual data collections. These catalogs offer definitions of common linguistic constructs (in the sense in which the term *construct* is used in psychology, see Cronbach and Meehl 1955) or *comparative concepts* in the sense of Haspelmath (2010). The structure of these metadata collections is crucially dependent on their general nature, and it is therefore not possible to provide a definition of this structure beforehand. However, what is possible is to say that reference catalogs tend to provide an *identifier* for a given construct, such as, a given *language* or a given *concept* or a given *speech sound*, along with (a) datasets that might make use of this construct, and (b) references that might define this construct. As a result, a reference catalog links an identifier to additional resources, which may refer to some literature references (typically modeled in BibTeX) or to some additional datasets (which could also be referred to with the help of URLs or DOIs). Major reference catalogs used in CLDF are (a) Glottolog, the reference catalog for language identifiers (Hammarström et al. 2021), (b) Concepticon, a reference catalog for concepts (List et al. 2022b), and (c) CLTS, a reference catalog for speech sounds. The major advantage of these reference catalogs is that they outsource the business of providing a consistent standard. Linguists making use of them do no longer need to define the constructs themselves, instead, they can link – where available – to the reference catalogs which take care of "the rest", by providing additional resources and by also taking the blame if the information they offer is wrong.

Why is it important to model concepts as constructs in linguistic research?

2 Glottolog

Glottolog (`https://glottolog.com`, Hammarström et al. 2021) is a reference catalog for language varieties and offers not only the identifiers for more than 7000 language varieties, but also an extensive bibliography that characterizes these language varieties. In this form, Glottolog is an excellent starting point for those who want to learn more about a particular language variety, since alone the bibliography delivers almost exhaustively all information that is available for individual languages.

In addition, Glottolog offers a preliminary classification of the language varieties in the form of language trees. This classification is close to the communis opinio in the field, but given that there are many different opinions here, no phylogeny can ever be perfect, and one should rather use the phylogeny offered by Glottolog as a convenient reference, rather than ground truth.

Glottolog also offers geolocations for most of the varieties the catalog contains. This is extremely convenient, since it means one can plot languages easily on a map, when having obtained their *Glottocodes*, the unique identifiers offered by the reference catalog, consisting of four letters and four numbers, derived based on the following criteria, outlined in Forkel and Hammarström (2022: 918)

- An ID specifically designed for machine readability, not confusable with an informal or human-directed identifier
- An ID type oblivious to level of linguistic abstraction (idiolect, sociolect, dialect, language, subfamily, family, etc.)

- An ID system for languages that improves on the ISO 639-3 language identifiers in terms of quality, transparency and anchoring

In addition, Glottolog can be accessed through a powerful Python API that offers users the possibility to search for language varieties, to extract trees in standardized formats, and to query all information also displayed on the website.

What is the difference between a language and a language variety?

3 Concepticon

Concepticon (List et al. 2016, List et al. 2022b) ist ein Katalog von sogenannten Concept Sets, einer Verlinkung von Questionnaires, wie Swadesh-Listen, etc., die für inzwischen mehr als 2900 Konzepte Definitionen und Links zu existierenden Questionnaires liefert. Das Concepticon ist essentiell für die Aggregierung von Daten, aber auch aus historischer Perspektive interessant. Erhältlich ist das Concepticon unter `http://concepticon.clld.org`. Wir werden uns in einer Sitzung den theoretischen Grundlagen des Concepticons widmen und in einer weiteren Sitzung lernen, wie wir die Software verwenden können.

Kann man Konzepte überhaupt definieren?

Background

In 1950, Morris Swadesh (1909 – 1967) proposed the idea that certain parts of the lexicon of human languages are universal, stable over time, and rather resistant to borrowing. As a result, he claimed that this part of the lexicon, which was later called *basic vocabulary*, would be very useful to address the problem of subgrouping in historical linguistics:

> [...] it is a well known fact that certain types of morphemes are relatively stable. Pronouns and numerals, for example, are occasionally replaced either by other forms from the same language or by borrowed elements, but such replacement is rare. The same is more or less true of other everyday expressions connected with concepts and experiences common to all human groups or to the groups living in a given part of the world during a given epoch. (Swadesh 1950: 157)

He illustrated this by proposing a first *list of basic concepts*, which was, in fact, nothing else than a collection of concept labels, as shown below:[1]

> I, thou, he, we, ye, one, two, three, four, five, six, seven, eight, nine, ten, hundred, all, animal, ashes, back, bad, bark, belly, big, [...] this, tongue, tooth, tree, warm, water, what, where, white, who, wife, wind, woman, year, yellow. (ibid.: 161)

In the following years, Swadesh refined his original concept lists of basic vocabulary items, thereby reducing the original test list of 215 items first to 200 (Swadesh 1952) and then to 100 items (Swadesh 1955). Scholars working on different language families and different datasets provided further modifications, be it that the concepts which Swadesh had proposed were lacking proper translational equivalents in the languages they were working on, or that they turned out to be not as stable and universal as Swadesh had claimed (Alpher and Nash 1999, Matisoff 1978). Up to today, dozens of different concept lists have been compiled for various purposes.

Who was one of the earliest Chinese scholars to propose a specific concept list?

[1] This list contains 123 items in total. According to Swadesh, these items occurred both in his original test list of English items, and in the data on the Salishan languages, which he employed for his first glottochronological study.

Concept Lists

Concept lists are simply speaking collections of concepts which scholars decided to compile at some point. In an ideal concept list, concepts would be described by a *concept label* and a short *definition*. Most published concept lists, however, only contain a concept label. On the other hand, certain concept lists have been further expanded by adding structure, such as *rankings*, *divisions*, or *relations*.

Concept lists are compiled for a variety of different *purposes*. The purpose for which a given concept list was originally defined has an immediate influence on its *structure*. Given the multitude of use cases in both synchronic and diachronic linguistics, it is difficult to give an exhaustive and unique classification scheme for all concept lists which have been compiled in the past. In the following table, we have nevertheless tried to distinguish eight basic types of concept lists and give one list for each of the types as a prototypical example.[2]

Type	Example	Purpose
basic vocabulary list ("Swadesh list")	Swadesh 1952 / 200 items	subgrouping
subdivided concept list	Yakhontov 1991 (Starostin 1991) / 35 + 65 items	genetic relationship, layer identification
"ultra-stable" concept list	Dolgopolsky 1964 / 15 items	genetic relationship
questionnaire	Allen 2007 / 500 items	dialect / language comparison
ranked list	Starostin 2007 / 110 items	subgrouping, layer identification
list of concept relations	DatSemShift, Bulakh et al. 2013 / 2424 items	representation of concept relations
special-purpose concept list	Matisoff 1978 / 200 items	subgrouping of Tibeto-Burman languages
historical concept list	Leibniz 1768 / 128 items	language comparison

Table 2: Examples for different types of concept list as they can be found in the literature

Linking Concept Lists

While all the concept lists which have been published so far constitute language resources with rich and valuable information, we lack guidelines, standards, best practices, and models to handle their interoperability. Language diversity is often addressed with region- or language-specific questionnaires. This makes it difficult to integrate and compare these resources.

The Concepticon is an attempt to overcome these difficulties by linking the many different concept lists which are used in the linguistic literature. In order to do so, we offer open, linked, and shared data in collaborative architectures. Our data is curated openly on GitHub (`https://github.com/clld/concepticon-data`). The Concepticon itself is published as Linked Open Data (`http://concepticon.clld.org`) within the CLLD framework, which allows us to reuse tools built on top of the CLLD API, in particular the `clldclient` package (`https://github.com/clld/clldclient`).

In our Concepticon, all entries from concept lists are partitioned into sets of labels referring to the same concept – so called *concept sets*. Each concept set is given a unique identifier (Concepticon ID), a unique label (Concepticon Gloss), a human-readable definition (Concepticon Definition), a rough semantic field, and a short description regarding its *ontological category*. Based on the availability of resources, we further provide metadata for concept sets, including links to the Princeton WordNet (University 2010), OmegaWiki (OmegaWiki 2005) and BabelNet (Navigli and Ponzetto 2012), and links to norm data bases, like SimLex-999 (Hill et al. 2015), the MRC Psycholinguistic database (Wilson 1988), and the Edinburgh Associative Thesaurus (Kiss et al. 1973).

[2] For further information regarding these concept lists, just click on the links in the "Example" field of the table.

A concept list is a collection of concepts that is deemed interesting by scholars. Minimally, it consists of an *identifier* for each concept which the lists contains, and a *label* by which the concept is referenced. The creator of a concept list is called a *compiler*. Each concept list is tied to one or more *sources*, it is given in one or more *source languages* and was compiled for one or more *target languages*. A *description* gives further information on each concept list in human-readable form, and tags are used to provide information regarding some basic characteristics of the concept list. The following figure illustrates how concept hierarchies are superimposed on our concept sets.

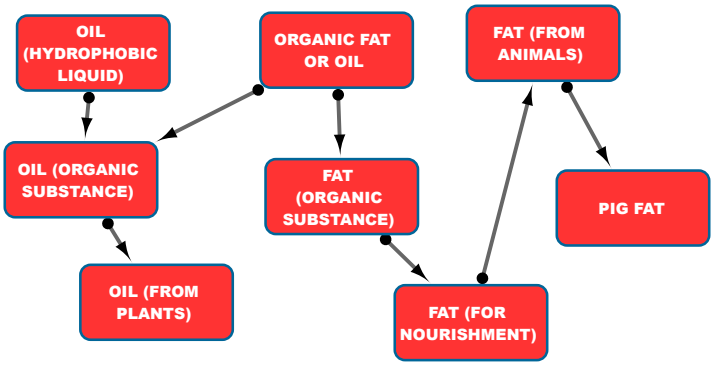

Figure 1: Concept relations between 'oil', and 'fat'

What is the concept from the semantic field for "fat" which we would expect in a Chinese questionnaire?

Examples

As a simple example for typical problems involving the linking of concept lists, consider the concepts given in the table below. Here, the four lists apparently intend to denote the same concept 'dull'. From the Chinese terms used in the lists by Ben Hamed and Wang (2006) and Chén (1996), however, we can clearly see that the intended meaning is not 'dull' in the sense of 'being blunt (of a knife)', but 'stupid'. Given that both authors originally wanted to render Swadesh's original concept lists in their research, this shows that we are dealing with a translation error here which may well result from the fact that in many concept lists, only 'dull' is used as a concept label, without further specification.

Compiler	Label	Concepticon
Blust (2008)	dull, blunt	DULL
Chén (1996)	呆，笨 / dull	STUPID
Comrie & Smith (1977)	dull	DULL
Wang (2006)	笨（ 不聪明 ）/ dull	STUPID
Swadesh 1952	dull (knife)	DULL

Table 3: Erroneous translations in concept lists

What other errors in translations can be possible, when considering Swadesh's original list of 200 concepts?

4 Cross-Linguistic Transcription Systems (CLTS)

CLTS (List et al. (2021), `https://clts.clld.org`) is a reference catalog for speech sounds, offering more than 8000 consistently defined speech sounds, which link to several databases. CLTS

is our standard for phonetic transcriptions in CLDF, specifically in the Lexibank collection (List et al. 2022a). It comes along with a rather powerful Python API that allows to conduct several operations with sounds, comparing sounds for their similarity based on distinctive features, or offering strategies to translate the representation of sounds across different transcription systems.

How is it possible that there are more than 8000 different sounds in human languages?

Background

Many linguists think that the International Phonetic Alphabet as defined by the International Phonetic Association is a clear-cut standard that does not leave any doubt and just has to be taken seriously by linguists (IPA 1999). However, if we look at the ways in which linguists produce linguistic data, we can first see, that the IPA is not the only phonetic transcription system currently in use. In addition, there is also the *North American Phonetic Alphabet* which is inconsistently and differently used by authors working chiefly on North American languages. There is the *Uralic Phonetic Alphabet*, which is often used but has also never been rigorously standardized (Sovijärvi and Peltola 1970). There is the *Lautschrift der Theutonista* (Wiesinger 1964) which was chiefly used to transcribe German dialect varieties, and there are the specific but largely regular idiosyncrasies of Chinese dialectologists who still keep using an older IPA version from the 1970ies.

Does it really make a difference, which transcription systems linguists use?

Problems

As a result of this high number of different transcription systems, we encounter many problems when trying to make our data cross-linguistically comparable. Essentially, if linguists say that their data has “IPA inside” this may mean different things depending on the linguists. In addition, the IPA itself creates ambiguities and does not consider itself as a standard in the common sense, but more as a set of suggestions that should help linguists carrying out phonetic transcriptions. Unfortunately, linguists even disregard the suggestions made by the IPA, not to speak of many pitfalls resulting from the Unicode standard and its use (Moran and Cysouw 2018).

Why does the IPA not want to be a standard?

Comparative Databases

As of now, there are many comparative databases which offer interesting cross-linguistic data, mainly for phoneme inventories in the languages of the world, but sometimes even containing lexical descriptions. The following table gives an overview on some larger datasets:

Dataset	Transcr. Syst.	Sounds
GLD (Ruhlen 2008)	NAPA (modified)	600+ (?)
Phoible (Moran et al. 2019)	IPA (specified)	2000+
GLD (Starostin 2015)	UTS	?
ASJP (Wichmann et al. 2016)	ASJP Code	700+
PBase (Mielke 2008)	IPA (specified)	1000+
Wikipedia	IPA (unspecified)	?
JIPA	IPA (norm?)	800+

Table 4: Cross-linguistic datasets with different transcription systems

What is the JIPA?

Objective of CLTS

The goal of CLTS is to provide a standard for phonetic transcription for the purpose of cross-linguistic studies by offering standardized ways to represent sound values serve as ”comparative concepts” in the sense of Haspelmath (2010). Similar to the Concepticon, we want to allow to register different transcription systems but link them with each other by linking each transcription system to unique sound segments. In contrast to Phoible or other databases which list solely the inventories of languages, CLTS is supposed to serve as a standard for the handling of lexical data in the CLDF framework, as a result, not only sound segments need to be included in the framework, but also ways to transcribe lexical data consistently.

What consequences does it have if CLTS is supposed to serve for phonetic transcription of lexical entries?

Strategy

We register transcription systems by linking the sounds to phonetic feature bundles which serve as identifiers for sound segments. When being given a form that is supposed to be presented in a given transcription system, we apply a three-step normalization procedure that goes from (1) NFD-normalization (Unicode decomposed characters), via (2) Unicode confusables normalization (`http://unicode.org/cldr/utility/confusables.jsp`), to (3) dedicated *alias symbols*. We divide sounds in different sound classes (currently *vowel*, *consonant*, *diphthong*, *cluster*, *click*, *tone*) to define specific rules for their respective feature sets. Additionally, we allow for a quick expansion of the set of features and the sound segments for each alphabet by applying a procedure that tries to guess unknown sounds by decomposing them into base sounds and diacritics.

On top of the different sounds we can register in this way, we link the feature bundles with datasets, like Phoible, LingPy’s sound class system, Wikipedia’s sound descriptions, or the binary feature systems published along with PBase (see above for references). Our feature system is not ambitious, as it is neither minimal, nor ordered, nor exclusive, nor binary, as in features systems that have been proposed in the past (Chomsky and Halle 1968). They merely serve as a means of description, following the IPA as closely as possible. The following two tables illustrate how characters are analysed in CLTS.

Input	NFD	Confus.	Alias	Out
ã (U+00E3)	a (U+0061) ◌̃ (U+0303)			ã
a (U+0061) : (U+003a)		a (U+0061) ː (U+02d0)		aː
ʦ (U+02a6)		t (U+0074) s (U+0073)		ʦ

Table 5: Three-step normalization in CLTS.

Sound	Identifier
ã	nasalized unrounded open front vowel
aː	long unrounded open front vowel
ʦ	voiceless alveolar affricate consonant

Table 6: Identifiers for sounds.

Wouldn’t it be sufficient to go for simple NFD normalization, given that Unicode is a real standard?

API, Online Demo, and Statistics

The API is similar to the one which is shipped with the Concepticon and offers easy ways for experienced Python users to use the data for automatic analyses. In addition, we are working on an online demo, which currently exists as a prototype and can be accessed via `http://calc.digling.org/clts/`.

Our current statistics are constantly changing in this stage, and we expect to expand the data quickly during the next months. Currently, we have registered two transcription systems, B(road)IPA and ASJP, as well as two meta-data-sets (Phoible and PBase). The following table shows, how many sounds of Phoible and Pbase we already cover:

Dataset	Matched	Generated	Missed	Perc.
Phoible	613	616	772	61%
PBase	496	265	521	59%

Table 7: Current coverage of CLTS

What problems can be expected when trying to link all of the sounds in Phoible and Pbase?

Outlook

In the future, we plan to add four more transcription systems (UPA, NAPA, GLD-UTS, X-SAMPA), more more metadata (Index Diachronica, Ruhlen's Database, sound examples, examples from the JIPA), we want to enhance the Python API to work on all platforms, and all Python versions (2 and 3), and we want to enhance the web-application (allow to select between different transcription systems, translate between systems, etc.).

All nice, but what do you think can be done with all this "normalized" data? Why do we even need unified transcription systems?

5 Norms, Ratings, and Relations

A very recent reference catalog, called NoRaRe, the Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (`https://norare.clld.org`, Tjuka et al. 2022) collects conceptual metadata from psycholinguistic datasets for individual words and concepts across different languages. The major idea of the NoRaRe collection was to offer a way to consistently compare conceptual metadata collected in the context of psychology and psycholinguistics, but also in the context of computational linguistics across datasets. As of now, NoRaRe offers data for 113 datasets, from which 713 variables are derived. These variables can often be compared across languages. Thus, one can find frequency information not only for Spanish, but also for English, German, etc. The underlying concepts, for which these variables are defined, are consecutively linked to the Concepticon. As a result, data that links to Concepticon can also make active use of the norms, ratings, and relations offered in NoRaRe.

There is not a clear-cut distinction between words and concepts in the NoRaRe database, where words are often thought of as representing individual concepts. Is this handling of the data a problem, or can it be justified in some way?

6 Future Reference Catalogs

More reference catalogs may be produced in the future. Since the creation of reference catalogs is tedious, however, it is hard to tell what reference catalog will come next. Candidates are a reference catalog for the glosses used in interlinear-glossed text (e.g., in the form of a Grammaticon), or senses that are used to define morphemes in words (some kind of a Morphemicon), or an extended collection of metadata for individual speech sounds (some Phoneticon).

Why is there not yet a reference catalog for bibliographic entries?

References

Allen, B. (2007). *Bai Dialect Survey*. Dallas: SIL International. PDF: http://www.sil.org/silesr/2007/silesr2007-012.pdf.

Alpher, B. and D. Nash (1999). "Lexical replacement and cognate equilibrium in Australia." *Australian Journal of Linguistics: Journal of the Australian Linguistic Society* 19.1, 5–56.

Ben Hamed, M. and F. Wang (2006). "Stuck in the forest: Trees, networks and Chinese dialects." *Diachronica* 23, 29–60.

Bulakh, M., D. Ganenkov, I. Gruntov, T. Maisak, M. Rousseau, and A. Zalizniak, eds. (2013). *Database of semantic shifts in the languages of the world*. URL: http://semshifts.iling-ran.ru/ (visited on 11/04/2014).

Chén, B. 陈保亚. (1996). *Lùn yǔyán jiēchù yǔ yǔyán liánméng* 论语言接触与语言联盟 *[Language contact and language unions]* [Language contact and language unions]. Běijīng 北京: Yǔwén □□.

Chomsky, N. and M. Halle (1968). *The sound pattern of English*. New York, Evanston, and London: Harper and Row.

Cronbach, L. J. and P. E. Meehl (1955). "Construct validity in psychological tests." *Psychological Bulletin* 52, 281–302.

Dolgopolsky, A. B. "Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concering the oldest relationships among the language families of Northern Eurasia]." *Voprosy Jazykoznanija* 2 (1964), 53–63; English translation: Dolgopolsky, A. B. "A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia." In: *Typology, relationship and time. A collection of papers on language change and relationship by Soviet linguists. Typology, Relationship and Time. A collection of papers on language change and relationship by Soviet linguists*. Ed. and trans. from the Russian by V. V. Shevoroshkin. Ann Arbor: Karoma Publisher, 1986, 27–50.

Forkel, R. and H. Hammarström (2022). "Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information." *Semantic Web* 13.6. Ed. by J. Bosque-Gil, M. Dojchinovski, P. Cimiano, J. Bosque-Gil, P. Cimiano, and M. Dojchinovski, 917–924.

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.

Haspelmath, M. (2010). "Comparative concepts and descriptive categories." *Language* 86.3, 663–687.

Hill, F., R. Reichart, and A. Korhonen (2015). "SimLex-999: Evaluating semantic models with (genuine) similarity estimation." *Computational Linguistics* 41.4, 665–695.

IPA, ed. (1999). *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press.

Kiss, G., C. Armstrong, R. Milroy, and J. Piper (1973). "An associative thesaurus of English and its computer analysis." In: *The computer and literary studies*. Ed. by A. Aitken, R. Bailey, and N. Hamilton-Smith. Edinburgh: Edinburgh University Press, 153–165.

Leibniz, G. W. von (1768). "Desiderata circa linguas populorum, ad Dn. Podesta [Desiderata regarding the languages of the world]." In: *Godefridi Guilielmi Leibnitii opera omnia, nunc primum collecta, in classes distributa, praefationibus et indicibus exornata* [Collected works of Gottfried Wilhelm Leibniz, now first collected, divided in classes, and enriched by introductions and indices]. Ed. by L. Dutens. Vol. 6. 2. Geneva: Fratres des Tournes, 228–231.

List, J.-M., C. Anderson, T. Tresoldi, and R. Forkel (2021). *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., 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/.

Matisoff, J. A. (1978). *Variational semantics in Tibeto-Burman. The "organic" approach to linguistic comparison*. Philadelphia: Institute for the Study of Human Issues.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford: Oxford University Press.

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.

Navigli, R. and S. P. Ponzetto (2012). "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." *Artificial Intelligence* 193, 217–250.

OmegaWiki (2005). *OmegaWiki: A dictionary in all languages*.

Ruhlen, M. (2008). *A global linguistic database*. Moscow: RGGU.

Sovijärvi, A. and R. Peltola (1970). *Suomalais-Ulgrilainen Tarkekirjoitus* UralicPhoneticAlphabet. Transcription System. Helsinki: University of Helsinki.

Starostin, G. S. and P. Krylov, eds. (2011). *The Global Lexicostatistical Database. Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-form*. URL: http://starling.rinet.ru/new100/main.htm.

Starostin, S. A. (1991). *Altajskaja problema i proischoždenije japonskogo jazyka [The Altaic problem and the origin of the Japanese language]*. Moscow: Nauka.

Swadesh, M. (1950). "Salish internal relationships." *International Journal of American Linguistics* 16.4, 157–167. JSTOR: 1262898.

— (1952). "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos." *Proceedings of the American Philosophical Society* 96.4, 452–463.

— (1955). "Towards greater accuracy in lexicostatistic dating." *International Journal of American Linguistics* 21.2, 121–137. JSTOR: 1263939.

Tjuka, A., R. Forkel, and J.-M. List (2022). "Linking norms, ratings, and relations of words and concepts across multiple language varieties." *Behavior Research Methods* 54.2, 864–884.

University, P. (2010). *WordNet. A lexical database for English*. Online Resource. Princeton.

Wichmann, S., E. W. Holman, and C. H. Brown (2016). *The ASJP database*. Jena: Max Planck Institute for the Science of Human History.

Wiesinger, P. (1964). "Das phonetische Transkriptionssystem der Zeitschrift "Teuthonista". Eine Studie zu seiner Entstehung und Anwendbarkeit in der deutschen Dialektologie mit einem Überblick über die Geschichte der phonetischen Transkription im Deutschen bis 1924." *Zeitschrift für Mundartforschung* 31.1, 1–20.

Wilson, M. D. (1988). "The MRC psycholinguistic database: Machine readable dictionary. Version 2." *Behavioural Research Methods, Instruments and Computers* 20.1, 6–11.