Concepticon: A Resource for the Linking of Concept Lists

Johann-Mattis List[1], **Michael Cysouw**[2], **Robert Forkel**[3]

[1]CRLAO/EHESS and AIRE/UPMC, Paris, [2]Forschungszentrum Deutscher Sprachatlas, Marburg,
[3]Max Planck Institute for the Science of Human History, Jena
mattis.list@lingpy.org, cysouw@uni-marburg.de, forkel@shh.mpg.de

Abstract

We present an attempt to link the large amount of different concept lists (aka "Swadesh lists") which are used in the linguistic literature. This resource, the *Concepticon* (`http://concepticon.clld.org`), links 20 077 concept labels from 100 conceptlists to 2435 concept sets. Each concept set is given a unique identifier, a unique label, and a human-readable definition. Concept sets are further structured by defining different relations between the concepts. The resource can be used for various purposes. Serving as a rich reference for new and existing databases in diachronic and synchronic linguistics, it allows researchers a quick access to studies on semantic change, cross-linguistic polysemies, and semantic associations.
Keywords: concepts, concept lists, Swadesh lists, cross-linguistically linked data

1. Introduction

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 these parts 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, black, blood, bone, brother (elder), child (son or daughter), cloud, cold, come, cry (weep), dance, day, dog, dust, ear, earth, eat, egg, eye, far, father, fire, flower, fog, foot, good, grass, green, hair, hand, head, heart, here, hit (with fist), hunt, husband, ice, lake, laugh, leaf, left hand, leg, liver, long, louse, man, meat, mother, mountain, mouth, name, near, neck, night, nose, person, rain, red, right hand, road (trail), root, rope, salt, sand, short, sing, sister (elder), skin, sky, small, smoke, snake, snow, speak, spear (war), star, stone, sun, swim, tail, that, there, this, tongue, tooth, tree, warm, water, what, where, white, who, wife, wind, woman, year, yellow. (Swadesh, 1950, 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 (Matisoff, 1978; Alpher and Nash, 1999). Up to today, dozens of different concept list have been compiled for various purposes. They are are used as heuristical tools for the detection of deep genetic relationships among languages (Dolgopolsky, 1964), as basic values for traditional lexicostatistical and glottochronological studies (Dyen et al., 1992; Starostin, 1991), or as litmus test for dubious cases of language relationship which might be due to inheritance or borrowing (McMahon et al., 2005; Chén Bǎoyà 陈保亚, 1996; Wang, 2006).

Apart from concept lists proposed for the application in historical linguistics, there is a large amount of not explicitly diachronic data, including concept lists serving as the basis for field work in specific linguistic areas (Kraft, 1981), concept lists which serve as the basis for large surveys on specific linguistic phenomena (Haspelmath and Tadmor, 2009), or concept lists which deal with the internal *structuring* of concepts, be it cognitive associations (Nelson et al., 2004), cross-linguistic polysemies (List et al., 2014), or frequently recurring semantic shifts (Bulakh et al., 2013).

[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.

2. Concept Lists

Concept lists are – simply speaking – lists of concepts. In these lists, concepts are ideally 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*.

2.1. Purpose of Concept Lists

Concept lists are compiled for a variety of different purposes. A major distinction can be made between those concept lists which have been compiled for the purpose of *language comparison* and those which have been compiled for the purpose of *concept comparison*. Among the former, we can further distinguish those lists which are used to prove *language relationship* (Dolgopolsky, 1964), those which are used for *linguistic subgrouping* (Norman, 2003; Starostin, 1991; Swadesh, 1955), and those which can be used to identify *contact layers* (Chén Bǎoyà 陈保亚, 1996). Among the latter, we can distinguish between concept lists with a primarily *synchronic objective* (Hill et al., 2014), and those with a primarily *diachronic objective* (Haspelmath and Tadmor, 2009; Bulakh et al., 2013).

2.2. Structure of Concept Lists

The purpose for which a given concept list was originally defined has an intermediate 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 schema for all concept lists which have been compiled in the past. In Table 1, 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 (see 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 1: Examples for different types of concept list as they can be found in the literature.

3. 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. This is specifically important in the context of multilingual language resources and resources on less-well-studied languages. Language diversity is often addressed with region- or language-specific questionnaires. This makes it difficult to integrate and compare these resources on a greater scale. Despite the growing body, the interoperability of language resources involving concepts and meanings, like wordlists and lexical datasets, has not yet been addressed in a systematic way.

The Concepticon is an attempt to overcome these difficulties by linking the many different concept lists ("Swadesh Lists") which are used in the linguistic literature. In order to do so, we offer open, linked, and shared data and tools in open and collaborative architectures. Our data is curated openly and collaboratively 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 the Concepticon, all entries from concept lists are partitioned into sets of labels referring to the same concept – so called *concept sets*. The Concepticon currently[3] links 20 077 concepts from 100 concept lists to 2435 concept sets and defines 639relations between the concept sets.

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. The core data model of the Concepticon is illustrated in Figure 1.

To facilitate our workflow and to guarantee the comparability of concept lists even if they are not linked to a overlapping set of concept sets, we define additional and very simple *concept relations* between concept sets (*broader*, *narrower*, *similar*). Even if the concepts in two or more concept lists are not assigned to the same concept set, they can still be comparable if the respective concept sets are related.

4. Examples

Linking concept labels to concept sets may be simple and straightforward. It may, however, also turn into a rather complicated task for which no completely satisfying solution can be found. In the following, we illustrate the difficulties of linking concepts to concept sets

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

[3] This is version 0.2 "public beta", see `http://dx.doi.org/XXX`. Note that this version is not yet published online, but the data is already being distributed.

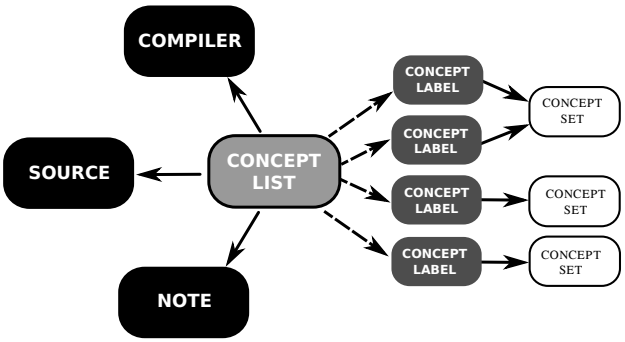

Figure 1: The core data model of the Concepticon.

with help of three seemingly less complicated examples. By this, we want to shed light on the theoretical and practical problems we had to face when compiling the Concepticon, and how we tried to address them.

4.1. CHILD: "Young Human" or "Descendant"?

As a first example for both the problems one faces when trying to link concepts across concept lists and the way we try to address them in the Concepticon, consider the different concept labels for "child" given in Table 2. As we can see from the labels themselves, the label "child" can denote two different concepts, of which one could be specified as "child (young human)" and the other as "child (descendant)". Not all concept lists, however, offer this precision. Swadesh himself, for example, would specify the "descendant" reading in his first list from 1950, but the "young human" reading in the list from 1952. In the list by Comrie and Smith (1977), which was intended to be a merger of the Swadesh's 200-item list from 1952 and his 100-item list from 1955, this specification is lost, and we cannot tell from the concept label which reading was intended by the compiler. The same applies for the lists of Blust (published in Greenhill et al., 2008) and Chén Bǎoyà 陈保亚 (1996). In order to handle these problems resulting from ambiguous concept labels, we assign those concepts whose reading we cannot determine from the concept label and the further descriptions given in the concept lists to a broader concept set "CHILD". Additionally, we set up a relation that states that "CHILD" is both broader as "CHILD (YOUNG HUMAN)" and "CHILD (DESCENDANT)". Figure 2 shows the relations involving "CHILD" which have been currently defined in the Concepticon.

4.2. RAIN: Thing or Action?

Another example for problems involving concept labels in concept lists are basic words related to "rain". Here, as illustrated in Table 3, the problem of mapping is not to find out which reading is intended, since "rain" itself is a rather clearcut concept, but it is not possible in all cases to tell whether the compilers intended to denote the *thing* or the *action*. This is a problem resulting from the use of English as a language for concept labels, since both the noun and the verb are homophones. In the lists of von Leibniz (1768) and Chén

Compiler	CONCEPT LABEL	Concepticon
Blust (2008)	child	CHILD
Chén (1996)	孩子/ child	CHILD
Comrie and Smith (1977)	child	CHILD
Leibniz (1768)	infans	CHILD (YOUNG HUMAN)
Matisoff (1978)	child/son	CHILD (DESCENDANT)
Swadesh (1950)	child (son or daughter)	CHILD (DESCENDANT)
Swadesh (1952)	child (young person rather than as relationship term)	CHILD (YOUNG HUMAN)

Table 2: Concept Labels and Concept Sets for "child".

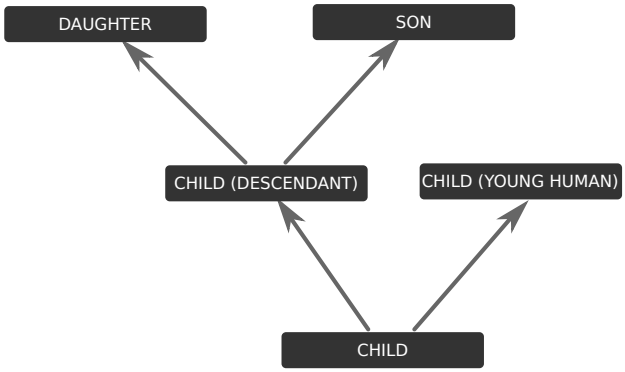

Figure 2: Concept Relations for "child".

Bǎoyà 陈保亚 (1996), there is no doubt that the *thing*-reading is intended, since noun and verb of "rain" are not homophone, neither in Chinese, nor in Latin. The same holds for the list by Blust (2008), since it structures the concepts into specific semantic fields which clearly indicate which reading is intended. In the lists of Swadesh (1950) and Comrie and Smith (1977), however, it is not possible to determine the intended reading. For this reason, we set up an overarching concept set "RAINING OR RAIN" which we define as being broader as "RAIN (PRECIPATION)" and "RAIN (RAINING)" (see Figure 3).

Compiler	CONCEPT LABEL	Concepticon
Blust 2008	rain	RAIN (PRECIPATION)
Chen 1996	雨/ rain	RAIN (PRECIPATION)
Comrie and Smith (1977)	rain	RAINING OR RAIN
Leibniz 1768	pluvia	RAIN (PRECIPATION)
Matisoff 1978	rain	RAIN (PRECIPATION)
Swadesh 1950	rain	RAINING OR RAIN
Swadesh 1952	to rain	RAINING

Table 3: Concept Labels for "rain"

4.3. "DULL" or "STUPID"?

As a last example for typical problems involving the linking of concept list, consider the concepts given in Table 4. Here, the four lists apparently intend to denote the same concept "dull". From the Chinese terms used in the lists by Wang (2006) and Chén Bǎoyà 陈

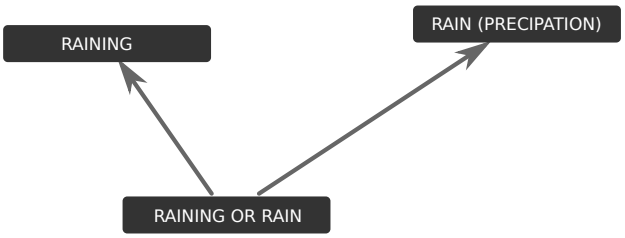

Figure 3: Concept Relations for “rain”.

保亚 (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	CONCEPT LABEL	Concepticon
Blust (2008)	dull, blunt	DULL
Chén (1996)	呆，笨/ dull	STUPID
Comrie and Smith (1977)	dull	DULL
Wang (2006)	笨（不聪明）/ dull	STUPID
Swadesh 1952	dull (knife)	DULL

Table 4: Erroneous Translations in Concept Lists.

5. Using the Concepticon

The Dictionaria project (`http://home.uni-leipzig.de/dictionaryjournal/`) provides a typical use case for the Concepticon: The project will publish dictionaries of minor languages in a way that maximizes opportunities for data reuse. In particular, comparability across dictionaries via *comparison meanings* (“standardized” meanings that ease the comparison) is a goal. Concepticon concept sets are the natural choice for such comparison meanings, and the concept labels linked to concept sets provide a good heuristic to match meaning descriptions of dictionary words to these sets. Since the Concepticon is open to extension, the Dictionaria project will also feed information back into the Concepticon by distilling and contributing a list of concepts commonly encountered in dictionaries of minor languages.

6. Outlook

An enormous amount of concept lists have been produced so far, not only in historical linguistics, but also in other disciplines that practically deal with the meanings of words, such as psycholinguistics, cognitive linguistics, but also second language learning. With the 100 concept lists we have assembled and linked in the Concepticon so far, we are still far away from getting anywhere near the top of the mountain. There are many more existing concept lists which need to be mapped to the Concepticon consecutively, and there is also important metadata, like BabelNet (`http://babelnet.org`) to which we want to link from our concept sets. Despite all the work that still needs to be done, we think that important first steps have been made with the Concepticon in its current form. In the future, we hope that we can advance further both by linking more lists to our resource in the future and by encouraging scholars to do the same with their data. With the collaborative efforts of the linguistic community, we may make an important step towards the standardization of concept lists.

7. Resource Information

The data underlying the Concepticon is curated at `http://github.com/clld/concepticon-data`. The application source code for the publication of the Concepticon can be accessed at `http://github.com/clld/concepticon`. The application itself can be accessed via `http://concepticon.clld.org`. The most recent official release of the Concepticon can be downloaded from `http://dx.doi.org/XXX`.

8. References

Allen, B. (2007). *Bai Dialect Survey*. SIL International.

Alpher, Barry and Nash, David. (1999). Lexical replacement and cognate equilibrium in australia. *Australian Journal of Linguistics: Journal of the Australian Linguistic Society*, 19(1):5--56.

Bulakh, M., Ganenkov, Dimitrij, Gruntov, Ilya, Maisak, T., Rousseau, Maxim, and Zalizniak, A. (2013). Database of semantic shifts in the languages of the world.

Chén Bǎoyà 陈保亚. (1996). *Lùn yǔyán jiēchù yǔ yǔyán liánméng*. Yǔwén 语文, Běijīng 北京.

Comrie, Bernard and Smith, Norval. (1977). Lingua descriptive series: Questionnaire. *Lingua*, 42:1--72.

Dolgopolsky, Aron B. (1964). Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. *Voprosy Jazykoznanija*, 2:53--63.

Dyen, Isidore, Kruskal, Joseph B., and Black, Paul. (1992). An indoeuropean classification. *Transactions of the American Philosophical Society*, 82(5):iii--132.

Greenhill, Simon J., Blust, Robert, and Gray, Russell D. (2008). The austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271--283.

Haspelmath, Martin and Tadmor, Uri. (2009). World loanword database.

Hill, Felix, Reichart, Roi, and Korhonen, Anna `. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Kraft, Charles H., editor. (1981). *Chadic wordlists*. Dietrich Reimer, Berlin.

List, J.-M., Mayer, T., Terhalle, A., and Urban, M. (2014). Clics: Database of Cross-Linguistic Colexifications.

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

McMahon, April, Heggarty, Paul, McMahon, Robert, and Slaska, Natalia. (2005). Swadesh sublists and the benefits of borrowing: An Andean case study. *Transactions of the Philological Society*, 103:147--170.

Nelson, Douglas L., McEvoy, Cathy, and Schreiber, Thomas A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behaviour Research Methods, Instruments, & Computers*, 36(4):402--407.

Norman, Jerry. (2003). The Chinese dialects. In Thurgood, G. and LaPolla, R., editors, *The Sino-Tibetan languages*, pages 72--83. Routledge, London and New York.

Starostin, Sergej Anatol'evic. (1991). *Altajskaja problema i proischoždenije japonskogo jazyka [The Altaic problem and the origin of the Japanese language]*. Nauka, Moscow.

Swadesh, Morris. (1950). Salish internal relationships. *International Journal of American Linguistics*, 16(4):157--167.

Swadesh, Morris. (1952). Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96(4):452--463.

Swadesh, Morris. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121--137.

von Leibniz, Gottfried Wilhem. (1768). Desiderata circa linguas populorum, ad dn. podesta. In Dutens, Louis, editor, *Godefridi Guilielmi Leibnitii opera omnia, nunc primum collecta, in classes distributa, praefationibus et indicibus exornata*, volume 6.2, pages 228--231. Fratres des Tournes, Geneva.

Wang, Feng. (2006). *Comparison of languages in contact. The distillation method and the case of Bai*. Institute of Linguistics Academia Sinica, Taipei.