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 which are used in the linguistic literature, ranging from Swadesh lists in historical linguistics to naming tests in clinical studies and psycholinguistics. This resource, our *Concepticon*, links 30 222 concept labels from 160 conceptlists to 2495 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 list, Swadesh list, naming test, word norms, 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 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. (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 and Wang, 2004).

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; Hill et al., 2014), cross-linguistic polysemies (List et al., 2014), or frequently recurring semantic shifts (Bulakh et al., 2013). Concept lists play also an important role in education, where they are used to measure and aid learners' progress (Dolch, 1936), in psycholinguistics, where different kinds of word norm data, like frequency and concreteness, are needed to control for variables in experiments (Wilson, 1988), and in public health studies, where standardized naming tests are used to assess the degree of aphasia or language disturbance (Nicholas et al., 1989; Ardila, 2007).

Given the multitude of concept data that has been published in the past, it is surprising that no real attempt has been carried out to provide reliable standards which would help scholars to compare concepts across resources or to define consistently which concepts they would use in a given study. Apparently available resources like the Princeton WordNet (Princeton University, 2010) or its counterparts in other languages are only partly useful for this purpose, since their synsets are supposed to reflect the concrete meaning

[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.

of words, and not the denotation range of concepts. When trying to link a given concept list to WordNet or BabelNet (Navigli and Ponzetto, 2012), like, for example, the famous 100-item list by Swadesh (1955), one meets quickly insurmountable obstacles, since the exact definition can often only be captured by using two or more WordNet synsets, not to speak even of the cases where no corresponding concept can be found.[2]

It is for this reason that we decided to build a new resource that links between published and popular concept lists from scratch. Our resource was explicitly designed to link existing concept lists and provide means to standardize future concept lists, and it is thus no lexical database like WordNet, also no multilingual database, like BabelNet, but an explicit meta-resource, in which we use *concept sets* to link concepts across different concept lists, creating a true *concepticon* in the spirit of Poornima and Good (2010).

2. 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*. 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).

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 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.[3]

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.

Our 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 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 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, following the semantic fields which are used in the World Loanword Database (Haspelmath and Tadmor, 2009), and a short description regarding its *ontological category* which correlates roughly with the conception of *parts of speech* when dealing with words in individual languages. Our Concepticon reflects the idea of *metaglosses* proposed by Cooper (2014), exceeding it in application range while falling back in the ontological grounding of concept sets, which we define and add on the basis of what we find in concept lists and what we deem to be important. Based on the availability of resources, we further provide metadata for concept sets, including links to the Princeton WordNet (Princeton 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).[4] The Concepticon currently[5] links 30 222 concepts from 160 concept lists to 2495 concept sets and defines 406 relations 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 *de-*

[2] Already seemingly simple cases like the concept which Swadesh labelled 'claw (nail)' are problematic, since of the seven synsets that WordNet gives for 'claw' and 'nail', none coincides with Swadesh's obvious intention to denote the keratin part at the legs of animals.

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

[4] We do not have full coverage for each of the resources, partially due to lacking data, partially since we have not yet had time to link all data. Our current coverage is: WordNet 53%, OmegaWiki 79%, BabelNet 35%, SimLex-999 18%, MRC 74%, and EAT 78%.

[5] This is version 1.0, see +++pending+++, published at `http://concepticon.clld.org`.

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

scription 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 core data model of our 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. These relations can get rather complex, yet they are important in order to guarantee that each concept label is only linked to one concept set. As an example, consider the concept that Swadesh (1955) labeled as ‘fat (grease)’. As Swadesh’s list from 1952 clearly shows, he was thinking of ‘fat’ as an organic substance, and the majority of concept lists follow this idea, although the labels are often varying. In South-East Asia, however, it is difficult to find a direct counterpart, which is why the concept is either narrowed down to ‘animal fat’ (Allen, 2007) or even ‘pig oil’ (Ben Hamed and Wang, 2006), or it is broadened to ‘fat/oil/grease’ (Sidwell, 2015). Figure 2 shows how we address these difficulties by defining additional relations between concepts denoting ‘oil’ and ‘fat’.

As another example that illustrates the complexity of concept relations in our Concepticon, Figure 3 shows our current network for kinship terms starting from ‘sibling’ (for more details on kinship terms, see Evans, 2011). Note that for all concepts in the network, we found concept lists in which the concepts were explicitly distinguished.

4. Examples

Linking concept labels to concept sets may seem to be simple and straightforward. In reality, however, it often turns into a rather complicated task for which no completely satisfying solution can be found. Furthermore, the task cannot be carried out in a fully automated fashion, for example, by automatically matching identical or similar concept labels across concept lists, since especially in lists of basic vocabulary there is large variation in labeling practice among authors. At times, this labeling practice can surface as two identical labels which mean very different things. As an example, consider the label ‘you’ which we find in the lists by Chén Bǎoyà 陈保亚 (1996) and Blust (Greenhill et al., 2008), but which is intended to denote ‘THOU’ (2nd person singular) in the first and ‘YE’ (2nd person plural) in the second case. In the former concept list, this becomes evident when consulting the Chinese gloss in the source. In the latter case, both ‘thou’ *and* ‘you’ occur, thus giving us the hint that ‘you’ is intended to point to the plural, not to the homophone singular form of the word. The case of ‘you’ and ‘thou’ is but one example, and there are numerous cases where it is only possible to decide what was originally intended when looking either at additional translations, or at the list as a whole.

In the following, we illustrate some typical difficulties one encounters when linking concepts to concept sets with three examples which seem to be rather unproblematic upon first sight, but turn out to be quite challenging upon deeper investigation. In this way, we want to shed light on the theoretical and practical problems we had to face when compiling our Concepticon, and how we tried to address them.

4.1. The ‘Child’: A Young Human or a Descendant?

As a first example, 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)’.

4.2. ‘Rain: A Thing or an 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 clear-cut concept, but it is not possible in all cases to tell whether

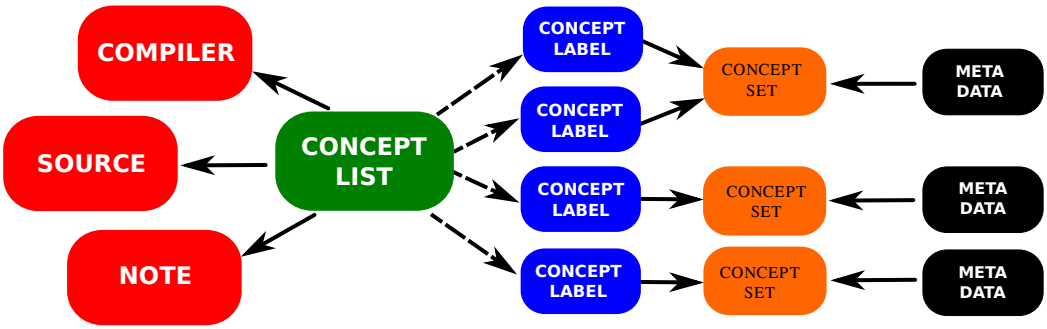

Figure 1: The core data model of our Concepticon. Concept lists are annotated by source, compiler ("creator"), and a human-readable note which provides further information. The concept lables ("glosses") in the concept lists are then assigned to concept sets. Concept sets themselves are further annotated by providing various kinds of meta data (links to WordNet, BabelNet, see text).

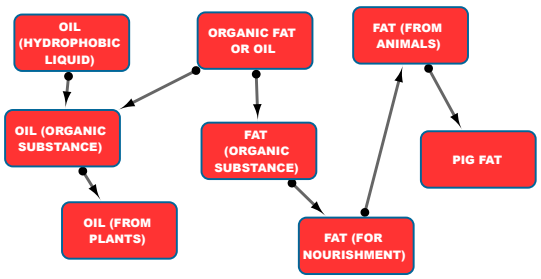

Figure 2: Concept relations between 'oil', and 'fat'

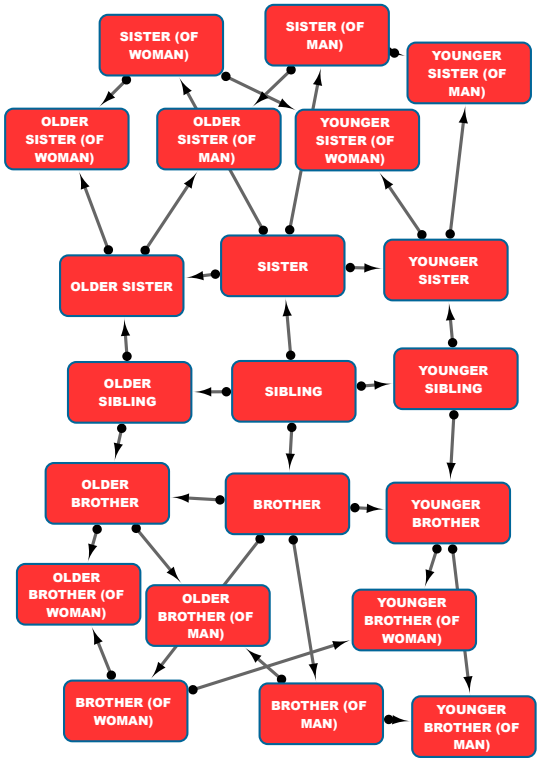

Figure 3: Concept relations between kinship terms

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 often homophones. In the lists of von Leibniz (1768)

Compiler	Label	Concepticon
Blust (2008)	child	CHILD
Chén (1996)	孩子/ child	CHILD
Comrie & Smith (1977)	child	CHILD
Dunn (2012)	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)
Tadmor (2009)	child (kin term)	CHILD (DESCENDANT)
Wiktionary (2003)	child (a youth)	CHILD (YOUNG HUMAN)

Table 2: Concept labels and concept sets for 'child'.

and Chén 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 (PRECIPITATION)' and 'RAIN (RAINING)'.

Compiler	Label	Concepticon
Blust 2008	rain	RAIN (PRECIPITATION)
Chen 1996	雨/ rain	RAIN (PRECIPITATION)
Comrie & Smith (1977)	rain	RAINING OR RAIN
Leibniz 1768	pluvia	RAIN (PRECIPITATION)
Matisoff 1978	rain	RAIN (PRECIITPATION)
Swadesh 1950	rain	RAINING OR RAIN
Swadesh 1952	to rain	RAIN (RAINING)

Table 3: Concept labels for 'rain'

4.3. 'Dull: Blunt 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 Ben Hamed and Wang (2006) and Chén Bǎoyà 陈保亚 (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	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 4: Erroneous translations in concept lists

5. Statistics

It is interesting to consider some basic statistics of the data we have collected in our Concepticon. By interesting statistics, we do not mean the basic numbers, like the number of concept sets, the number of unique concept labels, or the average size of concept lists, but questions which may give us a hint regarding the practice of concept list compilation, the preference of specific concepts in specific areas, or the general assumption of scholars regarding the importance or stability of certain concepts.

As a very simple but effective example, which we also use to check the correctness of our proposed links, we can, for example, measure the diversity of concept labeling, that is, the degree by which scholars differ in the use of concept labels which we link to the same concept sets. This degree of *diversity* in concept labelling is important to check how well we have succeeded in linking the data, since wrongly assigned links will also yield diverse concept labels. On the other hand, it reflects scholars' problems to denote concepts when compiling their concept lists. The ten most diverse concepts in our Concepticon are given in Table 5.

No.	Concept Set	Conc. Labels	Conc. Lists
1	THOU	32	111
2	FAT (ORGANIC SUBSTANCE)	31	82
3	EARTH (SOIL)	25	103
4	PERSON	25	109
5	HAIR	23	117
6	LIE (REST)	23	65
7	BE ALIVE	22	66
9	RIGHT	21	73
10	ROAD	20	67

Table 5: Most diverse concept labels in our Concepticon

Another potentially interesting measure is the frequency by which concepts are included in concept lists. Since our Concepticon contains many word lists which were compiled for the purpose of language comparison, the concept frequency reflects how important certain concepts are for comparative work in linguistics. The ten most frequently linked concept sets are given in Table 6. When comparing this list with the list in Table 5, it becomes obvious that frequency does not directly imply diversity, although it is clear that the more frequently a concept is included in concept labels, the higher are the chances that it is labeled differently. We can also see that the list of the most frequent concept sets roughly reflects those concepts which historical linguists usually think to be the most stable ones.

No.	Concept Set	Conc. Labels	Conc. Lists
1	TOOTH	17	131
2	WATER	10	130
3	EYE	12	129
4	TWO	14	127
5	FIRE	10	126
6	STAR	9	125
7	TONGUE	12	125
8	BLOOD	10	125
9	EAR	13	124
10	ONE	9	124

Table 6: Most frequent concepts in our Concepticon

In order to investigate this further and also to illustrate the usefulness of our Concepticon as a resource, we carried out a little experiment on concept coverage in those concept lists which are either ranked, thus providing us with scores regarding the supposed stability of concepts, and those lists which we tagged as *ultra-stable*, thus reflecting concept lists collections in which scholars reported what they thought to be the most slowly changing concepts, be it globally or area-specific. This selection criterion yielded a total of 27 different concept lists. In a first run, we trimmed all lists down to a comparable size. This was done by considering only the first 60 words of all ranked lists, while the usually shorter ultra-stable lists were kept as they were. This yielded a total number of 269 concept sets distributed across all 27 concept lists. This number is already quite surprising, given that the majority of the lists we considered was supposed to reflect items of high stability. In a second step, we ranked the concept sets according to frequency of occurrence across the 27 concept lists and used this rank to determine a list of the 40 most frequently recurring concepts. The first ten items of this list are largely identical with the general list of most frequent concepts given in Table 6, apart from 'I' (first person singular), which is in our short list but not in the list in Table 6, and 'ONE', which is not among the first ten items in our short list. In a third step we compared to which degree the 27 concept lists would overlap with this new concept list of 40 items which are most often thought to be stable.

The results of this experiment are given in form of a heatmap in Figure 4, in which the concept lists are ranked from top to bottom regarding their overlap with our base list, and the concept sets are ranked according to frequency

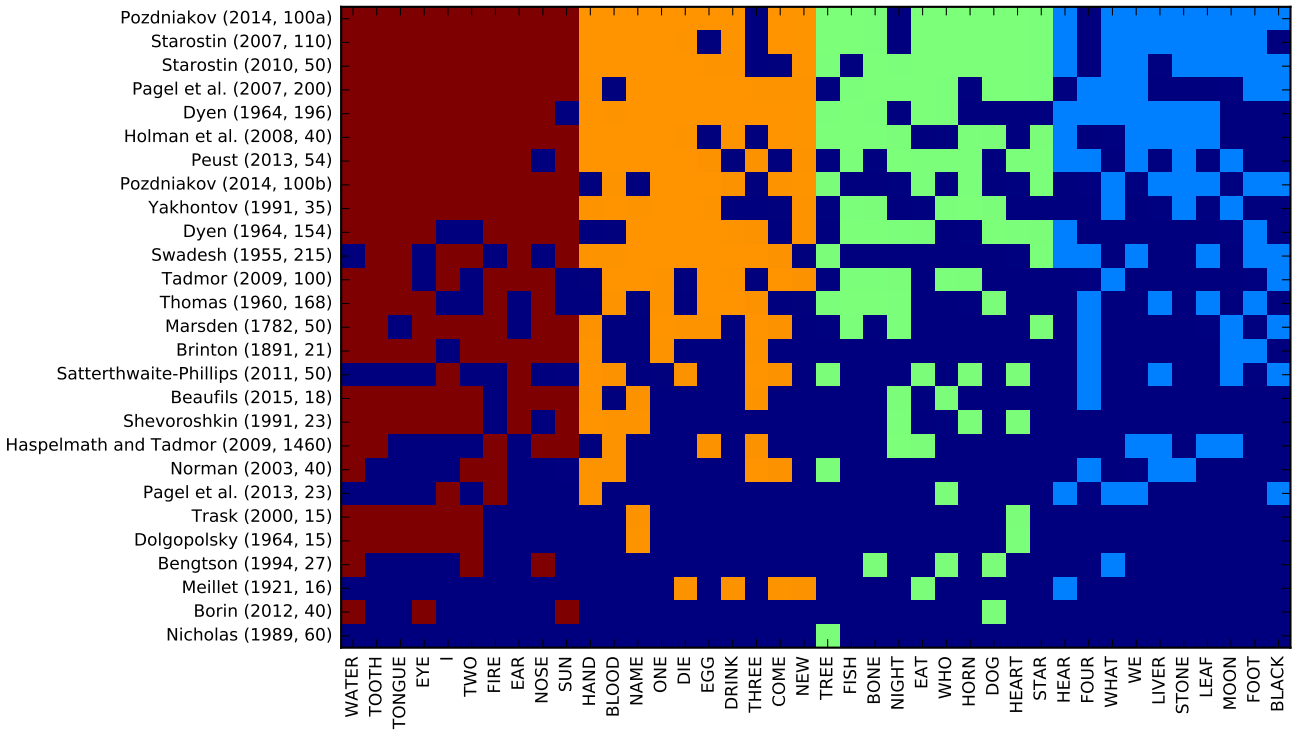

Figure 4: Comparison of stable words across ranked and ultra-stable concept lists. The figure shows a selection of 26 lists in our Concepticon which are tagged as either "ultra-stable" or "ranked". From ranked lists, the first 60 items were selected. From ultra-stable lists, all items were taken. The total number of concepts in these lists sums to 269, of which the first 40 were selected by taking those concepts which occurred across the largest number of concept lists. The figure lists the concepts in increasing rank from the left to the right. Blue spots in the row of a concept list indicate the absence of the item, colored spots (with stability ranks going from red to blue) indicate its presence.

of occurrence from left to right. Since this experiment reflects what scholars *assume* to be stable, and also due to general problems of defining *stability* in historical linguistics and typology (Dediu and Cysouw, 2013), the results should be taken with a certain care. The fact, however, that we find a rather large agreement in many lists whose stability scores were derived from quite different methods shows that it might be worthwhile to carry out a much closer comparison of different stability proposals than has previously been done. It is further clear that our Concepticon with all the already linked concept lists and the additional metadata provides the ideal starting point for deeper research on concept stability.

6. Using our Concepticon

The Dictionaria project (`http://home.uni-leipzig.de/dictionaryjournal/`) provides a typical use case for our Concepticon: The project will publish dictionaries of lesser-described 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.

7. 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 160 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, we have to expand the coverage of existing metadata, and there is also important metadata, like WikiData (`https://www.wikidata.org`) to which we want to link from our concept sets in the future. Despite all the work that still needs to be done, we think that important first steps have been made with our Concepticon in its current form. In the future, we hope that we can advance further both by linking more lists to our resource 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 concepts and concept lists.

Resource Information

The data underlying our Concepticon is curated at `http://github.com/clld/concepticon-data`. The appli-

cation 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 release of the Concepticon can be downloaded from +++pending+++.

Acknowledgements

This research was supported by the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history* (JML) and the ERC starting grant 240816 *Quantitative modelling of historical- comparative linguistics* (JML, MC). As part of the CLLD project (`http://clld.org`) and the GlottoBank project (`http://glottobank.org`), the work was supported by the Max Planck Society, the Max Planck Institute for the Science of Human History and the Royal Society of New Zealand (Marsden Fund grant 13-UOA-121, RF). All support is gratefully acknowledged. Many people helped us in many ways in assembling the data. They pointed us to missing lists (M), provided scans (S), digitized data (D), typed off and corrected concept lists (C), provided translations (T), linked concept lists (L), or gave important advice (A). For all this help, we are very grateful and express our gratitude to Alexei Kassian (D), Andrew Kitchen (D), Damian Satterthwaite-Phillips (D), Frederike Urke (CLS), Harald Hammarström (DMS), Julia Fischer (SDT), Lars Borin (DL), Martin Haspelmath (AD), Nicholas Evans (A), Paul Heggarty (D), Robert Blust (D), Sean Lee (D), Sebastian Nicolai (CMS), Thiago Chacon (MSD), and Viola Kirchhoff da Cruz (CLS).

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.

Ardila, Alfredo. (2007). Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology*, 22(3):297 -- 307. Special Issue: Cultural Diversity.

Beaufils, Vincent. (2015). eLinguistics.net. Quantifying the genetic proximity between languages. URL: `http://www.elinguistics.net`.

Ben Hamed, Mahe and Wang, Feng. (2006). Stuck in the forest: Trees, networks and chinese dialects. *Diachronica*, 23:29--60.

Bengtson, John D. and Ruhlen, Merritt. (1994). Global etymologies. In Ruhlen, Merritt, editor, *On the origin of languages: Studies in linguistic taxonomy*, pages 227--236. Stanford University Press, Stanford.

Borin, Lars. (2012). Core vocabulary: A useful but mystical concept in some kinds of linguistics. In Santos, Diana, Lindén, Krister, and Ng'ang'a, Wanjiku, editors, *Shall we play the Festschrift Game?*, pages 53--65. Springer, Berlin and Heidelberg.

Brinton, Daniel G. (1891). *The American race. A linguistic classification and ethnographic description of the native tribes of North and South America*. N. D. C. Hodges, New York.

Bulakh, M., Ganenkov, Dimitrij, Gruntov, Ilya, Maisak, T., Rousseau, Maxim, and Zalizniak, A. (2013). Database of semantic shifts in the languages of the world. URL: `http://semshifts.iling-ran.ru/`.

Chén Bǎoyà 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]. Yǔwén 语文, Běijīng 北京.

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

Cooper, Doug. (2014). Data Warehouse, Bronze, Gold, STEC, Software. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 91--99.

Dediu, Dan and Cysouw, Michael. (2013). Some structural aspects of language are more stable than others: A comparison of seven methods. *PLoS ONE*, 8(1):1--20, 01.

Dolch, E. W. (1936). A basic sight vocabulary. *The Elementary School Journal*, 36(6):456--460.

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.

Dyen, Isidore. (1964). On the validity of comparative lexicostatistics. In *Proceedings of the international congress of linguistics*, pages 238--252, Leiden. Sijthoff.

Evans, Nicholas. (2011). Semantic typology. In Sung, Jae Jung, editor, *The Oxford Handbook of linguistic typology*, pages 504--533. Oxford University Press, Oxford.

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*. Max Planck Digital Library, Munich.

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

Hill, Felix, Reichart, Roi, and Korhonen, Anna `. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4): 665--695.

Holman, Eric W., Wichmann, Søren, Brown, Cecil H., Velupillai, Viveka, Müller, André, and Bakker, Dik. (2008). Explorations in automated lexicostatistics. *Folia Linguistica*, 20(3):116--121.

Kiss, G., Armstron, Christine, Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. In Aitken, A.J., Bailey, R.W., and Hamilton-Smith, editors, *The computer and literary studies*, pages 153--165. Edinburgh University Press, Edinburgh.

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.

Marsden, William. (1782). XXI. Remarks on the Sumatran Languages. *Archaeologia*, 6:154--158.

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.

Meillet, A. (1965). *Linguistique historique et linguistique générale* [Historical and general linguistics]. Libr. Champion, Paris.

Navigli, Roberto and Ponzetto, Simone Paolo. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217--250.

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, and Computers*, 36(4):402--407.

Nicholas, Linda E., Brookshire, Robert H., MacLennan, Donald L., Schumacher, James G., and Porrazzo, Shirley A. (1989). The Boston Naming Test: Revised administration and scoring procedures and normative information for non-brain-damaged adults. In *Clinical Aphasiology Conference*, pages 103--115, Boston. College-Hill Press.

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.

OmegaWiki. (2005). OmegaWiki: A dictionary in all languages. URL: `http://www.omegawiki.org/`.

Pagel, Mark D., Atkinson, Quentin D., and Meade, Andrew. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163):717--721.

Pagel, Mark, Atkinson, Quentin D., Calude, Andreea S., and Meade, Andrew. (2013). Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Acadamy of Science, USA*, 110(21): 8471--8476.

Peust, Carsten. (2013). Towards establishing a new basic vocabulary list (Swadesh list). Draft.

Poornima, Shakthi and Good, Jeff. (2010). Modeling and encoding traditional wordlists for machine applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground.*, pages 1--9, Stroudsburg.

Pozdniakov, Konstantin. (2014). O poroge rodstva i indekse stabil'nosti v bazisnoj leksike pri massovom sravnenii: Atlantičeskie jazyki [On the threshold of relationship and the "stability index" of basic lexicon in mass comparison: Atlantic languages]. *Journal of Language Relationship*, 11:187--237.

Princeton University. (2010). Wordnet. URL `https://wordnet.princeton.edu/`.

Satterthwaite-Phillips, D. (2011). *Phylogenetic inference of the Tibeto-Burman languages*. Phd thesis, Stanford University, Stanford.

Shevoroshkin, V. V. and Manaster Ramer, A. (1991). Some recent work on the remote relations of languages. In Lamb, S. M. and Mitchell, E. D., editors, *Sprung from Some Common Source: Investigations into the Prehistory of Languages*, pages 178--199. Standorf University Press, Stanford.

Sidwell, Paul. (2015). Austroasiatic dataset for phylogenetic analysis: 2015 version. *Mon-Khmer Studies (Notes, Reviews, Data-Papers)*, 44:lxviii--ccclvii.

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

Starostin, S. A. (2007). Opredelenije ustojčivosti bazisnoj leksiki [determining the stability of basic words]. In *S. A. Starostin: Trudy po jazykoznaniju* [S. A. Starostin: Collected works on linguistics], pages 580--590. Languages of Slavic Cultures, Moscow.

Starostin, George S. (2010). Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship*, 3:79--116.

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.

Tadmor, Uri. (2009). Loanwords in the world's languages. Findings and results. In Haspelmath, Martin and Tadmor, Uri, editors, *Loanwords in the world's languages. A comparative handbook*, pages 55--75. de Gruyter, Berlin and New York.

Thomas, David D. T. (1960). Basic vocabulary in some Mon-Khmer languages. *Anthropological Linguistics*, 2(3):7--11.

Trask, Robert L. (2000). *The dictionary of historical and comparative linguistics*. Edinburgh University Press, Edinburgh.

von Leibniz, Gottfried Wilhem. (1768). Desiderata circa linguas populorum, ad Dn. Podesta [Desiderata regarding the languages of the world]. In Dutens, Louis, editor, *Godefridi Guilielmi Leibnitii opera omnia* [Collected works of Gottfried Wilhelm Leibniz], volume 6.2, pages 228--231. Fratres des Tournes, Geneva.

Wang, Feng and Wang, William S.-Y. (2004). Basic words and language evolution. *Language and Linguistics*, 5(3): 643--662.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary. Version 2. *Behavioural Research Methods, Instruments and Computers*, 20(1):6--11.