Towards a history of concept list compilation in historical linguistics

Johann-Mattis List

Max Planck Institute for the Science of Human History, Jena

A large proportion of lexical data of the world's languages is presented in form of word lists in which a set of concepts was translated into the language varieties of a specific language family or geographic region. The basis of these word lists are *concept lists*, that is, *questionnaires* of *comparative concepts* (in the sense of Haspelmath 2010), which scholars used to elicit the respective translations in their field work. Thus, a concept list is in the end not much more than a bunch of *elicitation glosses*, often (but not necessarily) based on English as an elicitation language, and a typical concept list may look like the following one, quoted from Swadesh (1950: 161):

> 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.

But scholars may also present their concept list in tabular form, adding additional information in additional columns, numbering and ranking items, providing exemplary translations into other languages, or marking specific items as obsolete.

The compilation of concept lists for the purpose of historical language comparison has a long tradition in historical linguistics, dating back at least to the 18th century (Leibniz 1768, Pallas 1786), if not even earlier (see Kaplan 2017). But concept lists were not solely compiled for the purpose of historical language comparison. If we employ the rough criterion by which any list of comparison concepts that was compiled for some scientific purpose can be seen as a concept list, we can find many more examples in the linguistic and scientific literature, including typological surveys (Brinton 1891), attempts to establish a language for global communication (Ogden 1930), or naming tests in clinical and psychological studies (Nicholas et al. 1989).

One of the most popular usage examples of concept lists in historical linguistics is Swadesh's theory of glottochronology (Swadesh 1952; Swadesh 1955), which stated that language splits can be dated due to the regular decay of words in the basic vocabulary of languages. Although this idea was heavily criticized soon after it was first proposed, even more concept lists have been compiled since then, and scholars have not given up the idea that a list of universal and stable concepts expressed in all languages of the world could indeed be found (Brown et al. 2008; Dolgopolsky 1964; Shevoroshkin and Manaster Ramer 1991).

If one is solely interested in the "surface history" of concept lists, the story is quickly told: at some point in the history of science, when the interest of scholars arose to compare the world's linguistic diversity more systematically, some scholars quickly realized that it might be useful to use a fixed list of concepts (represented by elicitation glosses for different meanings) to identify diverging

Please cite as: List, Johann-Mattis. 2018. Towards a history of concept list compilation in historical linguistics. *History and Philosophy of the Language Sciences*. http://hiphilangsci.net/2018/10/31/concept-list-compilation/

pronunciations or different expressions. One could further ask who was the first to propose this practice and why this practice became so successful, and along with our increasing knowledge about the history of science, the dating of the first concept list ever compiled would constantly be shifted back. A different, and — in my opinion — much more interesting perspective, however, would ask, how scholars influenced each other, from whom they borrowed their ideas when assembling their concept lists, and to which degree they tried to circumvent the numerous problems that we always face when dealing with semantics.

In the following, I will try to illustrate how at least some of these questions can be answered with help of the *Concepticon* resource (List et al. 2016), a collaboratively curated database that has the ambitious goal of making all concept lists that were compiled in the past comparable with each other. In order to do so, I will first introduce the Concepticon project in more detail. I will then present basic types of concept lists that have been compiled in the past. Finally, I will try to give some examples on how the Concepticon resource can be used to study the practice of concept list compilation from a "stemmatic" perspective, that seeks to reconstruct the evolution of ideas that accumulated in the huge diversity of different concept lists that we can find in the literature today.

The Concepticon Project

The Conception project (https://concepticon.clld.org) presents an attempt to link the large number of different concept lists which were and are used in the linguistic literature with each other, in order to make explicit which concept list employs the same concepts (despite different elicitation glosses). Many scholars have and had their personal mappings between different concept lists. These allow them to aggregate data from different sources or to compare their research with the research of others. The Concepticon tries to make these mappings explicit by linking the elicitation glosses used in the various concept lists compiled so far to uniquely identified *concept sets*.

Linking Elicitation Glosses to Concepticon Concept Sets

The basic idea is that a given Concepticon concept set provides concrete information as to which concept lists try to elicit identical concepts, and what elicitation glosses they use. The following table illustrates this, by contrasting concrete elicitation glosses that we could find in different concept lists for our Concepticon concept set DUST.

Elicitation Gloss	Elicitation Language	Concept List
灰塵	Chinese	Beijing University 1964
$dust	English	Bowern et al. 2011
DUST	English	Youn et al. 2016
pulvis	Latin	Pallas 1789
DUST, (ASH)	English	Payne 1991
пыль	Russian	Mennecier et al. 2016

The major difference in the elicitation glosses used by the authors here are the language they use for glossing, but the differences in glossing practice can be much more than this, especially when dealing with concepts like personal pronouns, where we encounter the highest diversity in glossing, as can be seen from the following table illustrating some glosses for the Concepticon concept set THOU.

Eliciation Gloss	**Elicitation Language**	**Concept List**
you (sg.)	English	Alpher and Nash 1999
You	English	Beaufils 2015
thou	English	Benedict 1976
you.SG	English	Bowern and Atkinson 2012
you	English	Bowern 2012
you (int. sing.)	English	Cross 1964
second person marker	English	Dolgopolsky 1964
thou (you sg.)	English	McMahon et al. 2005

While colleagues who hear the first time about the Concepticon often think that the differences in elicitation glosses are negligible, and that it would be easy to detect what gloss is intended to gloss what concept, especially the example of THOU should make clear that the ambiguities introduced by using English as a glossing language cannot be underestimated. If a list, published in some article, without further comment, only provides “you” as an elicitation gloss, we can simply not tell what the gloss was originally intended to elicit: the singular or the plural. Usually we can only derive the information from the list as a whole. If the list contains an item “you (pl.)” as well, we know that “you” points to the singular, and likewise, if we find “you (sg.)”, we know the “you” is the plural. When linking concept lists in our Concepticon project, we try to take the greatest care to resolve all possible ambiguities, and where we cannot resolve them, we often also decide to leave concepts unlinked.

To make more explicit what elicitation glosses a given concept set assembles, each Concepticon concept set is further given a rough definition and a rough gloss, which are, however, not necessarily binding, and should never be taken literally, when inspecting the Concepticon resource, since it is well possible that a given gloss or definition for a Concept set is not useful or not precise, while the link to the different elicitation glosses in the concept list itself is coherent. In addition, we also provide metadata for each concept set, which we derived from different norm databases, including, among others, links to WordNet (Princeton University 2010), links to norm databases providing information on (language-specific) age of acquisition, or association measures (Kiss et al. 1973).

Current Concepticon Statistics

What are our current statistics with the Concepticon project? While the Concepticon website may be outdated, as we are constantly improving it, but only releasing the most recent version once or

twice per year, you can find the most recent statistics on our GitHub Repository, which by now provides the following numbers:

- concept sets (used): 3094
- concept lists: 223
- concept labels: 41587
- concept labels (unique): 10084

Concept *labels* here refers to what I have been calling *elicitation glosses* so far, that is, the concrete glosses that scholars use to elicit a certain concept (be it during fieldwork, or by translating with the help of dictionaries). The numbers should speak for themselves: we currently *use* 3094 concept sets to link 41587 different elicitation glosses in 223 different concept lists, of which 10084 are *unique*, i.e., they *all* differ from each other. Given that there are no simple ways to compare one elicitation gloss with another one (as I have tried to illustrate above), this emphasizes the importance of the Concepticon, since it renders comparable what was not comparable before.

Scholarly Concerns with the Concepticon Project

Scholars who hear about the Concepticon project for the first time are occasionally quite dismissive of our attempts. Among the typical concerns expressed, they think (a) that our approach was culturally biased, (b) that it was useless, because semantics was too fuzzy, and (c) that the implementation was problematic. The first concern is easily ruled out when inspecting the huge variety of concept lists we have assembled so far. If one inspects the concept lists that were linked in our project, one can find not only a huge variety of different source languages that were used for elicitation, but also a huge variety of concept lists that were designed to study specific languages in specific areas of the world. Since it is our goal to link (ideally) all existing concept lists that have been published so far, the only cultural bias that could result from this enterprise was the bias that is already inherent in the field of historical and diversity linguistics, since it is not us who decides what concepts people should query in fieldwork or typological surveys.

The second concern expresses the general attitude in the field of linguistics to blacklist certain and never question them again. We know this from the debate about the origin of language which was officially labelled as a question not belonging to the field of linguistics in the *statuts* of the Société de Linguistique de Paris (Statuts 1866). We know this also from generative syntax, where linguistic performance of a language was conveniently shelved away, licensing linguists to study linguistics via introspection, rather than with empirical techniques. We know this as well from semantics, which is thought to be so complex and fuzzy that it could never be reliably studied cross-linguistically. Since the Concepticon project touches a problem that is often black-listed, namely the definition of concepts, scholars may react very harshly when hearing about our attempts for the first time. However, the question of whether we can ever consistently define a concept or not, is not the question we ask in our project. Our purpose is to increase the comparability of data produced by linguists, and since many linguists compare elicitation glosses that are used across the literature in practice, we are only trying to make this practice transparent. The fact that we can break down more than 10000 unique (!) elicitation glosses to some 3000 different concept sets should be enough of a proof of concept. Even if we will never be able to define concepts consistently, we find enough

regularity to compare elicitation glosses across concept lists, providing specifically *practical help* to people working in the field of diversity linguistics.

The last concern, the concern regarding our *implementation*, usually arises around the definitions we provide for the concept sets. Many scholars are not content with them, as they find them misleading. Suggestions include to replace our definitions by linking consistently to WordNet instead of our concept sets, to use "standard" semantic theory, like *Natural Semantic Metalanguage* (Goddard 2010), to provide consistent definitions, or to simply take much more care in this regard. The misunderstanding here is that the definitions are just a type of metadata, some service we provide in addition to linking elicitation glosses to concept sets, they are not the core of our approach. We simply lack the working power to go through the more than 3000 concept sets and check all definitions, and we assembled the definitions we use from the existing literature already, including specifically the definitions used for the WOLD (Haspelmath and Tadmor 2009), which was our starting point, when we began with this endeavor. WordNet itself disqualifies as a source for our mappings, since it is a *dictionary* of a single language (English), lacking many of the complex concept sets we link to. Natural Semantic Metalanguage, on the other hand, has never been sufficiently expanded to account for the huge amount of concepts we try to provide identifiers for. Scholars also misunderstand that the Conception projects is based on *collaborative efforts*, so if anybody is unhappy with what we do, they are cordially invited to join our project and refine things they deem to be in need of refinement. We have a very clear contribution policy, and we consistently list all people who have helped us in the part as contributors, and we invite substantial contributors to join our editorial board, which is re-assembled during each new release.

Short History of the Concepticon Project

While we cite the official launch of the Concepticon project with the publication by List et al. (2016), when we first launched the project as a CLLD application, the origins of the project go back to the times when I started my PhD in Düsseldorf in 2009. During this time, I began (as many scholars before) to make my own mappings of different concept lists, especially those published by Swadesh and by the Moscow School of Historical Linguistics (represented prominently by Sergej and George Starostin). When I worked as a post-doc in Marburg under Michael Cysouw, we realized that we had a similar interest, but that Michael had a clearer idea of what the theoretical background of the project was, namely to establish some kind of a *lexicon of concepts*, or a *Concepticon*, as it was called by Poornima and Good (2010). We united our efforts, and Michael hired students to help to expand the initial mappings I had made in my time as a PhD student. Later, in 2014, Robert Forkel saw our initial attempts, by then presented in a self-made web-application, and immediately saw the potential of the project to help in data aggregation. He introduced many new ideas for a more consistent handling of the concept sets, including a Python software package that we since then use to check the data automatically for consistency. In 2015, Robert launched the first Concepticon CLLD application, which was then officially released along with our 2016 paper.

Since then, the Concepticon project has been further expanded. We have further increased the number of Concept lists (from originally about 160 to now more than 220), and we have also taken the Concepticon as a basis for the data sharing principles proposed by the Cross-Linguistic Data

Formats initiative (https://cldf.clld.org). The CLLD initiative itself was initiated in 2014 by a group of scholars under the lead of Robert and the by then newly founded Department of Linguistic and Cultural Evolution of the Max Planck Institute for the Science of Human History, directed by Russell D. Gray, and has now, after four years of hard work, published the CLDF format specifications in a first version (Forkel et al. 2018), which will hopefully contribute a lot to rendering the data we use in diversity linguistics more comparable in the future.

Towards an Evolutionary History of Concept List Compilation

The Concepticon project was in the first instance initiated to serve practical purposes. Given the large number of different datasets published in the past, scholars would like to aggregate them, in order to allow for a more consistent comparison of words across different datasets. A recent example illustrating the usefulness of the Concepticon to help in this regard, is the new version of the Database of Cross-Linguistic Colexifications (CLICS, List et al. 2018), where data from 15 different sources was successfully aggregated by linking the concept lists underlying each dataset to the Concepticon.

But the Concepticon is more than a simple tool for data aggregation. When linking new concept lists to our resource, we also pay careful attention to the circumstances under which a concept list was originally compiled. Each concept list is therefore characterized by a small text, written in prose, summarizing what we know about its origin. Surprisingly, we often do not know much in this regard, since what scholars tell us when they publish a concept list is not necessarily much. Many concept lists, for example, are presented as a "Standard Swadesh List" in the literature, but when inspecting the elicitation glosses, it becomes immediately clear that the scholars do not faithfully list any of the early lists published by Swadesh (Swadesh 1952, Swadesh 1955). Instead, scholars may introduce new concepts, misinterpret the sources, or take the elicitation glosses from intermediate sources. When digging deeper in the history of individual concept lists, it is surprising how intertwined their history is, and how inaccurately scholars deal with the problem of denoting a comparative concept they use in their research.

In the following, I will try to illustrate how the Concepticon resource can be used to study the history of concept list compilation. My illustration will only be anecdotal, based on things I noticed when expanding the Concepticon resource, and it is likely that some of my interpretations will turn out to be wrong. I will start form discussing the types of concept lists that we have assembled so far in the Concepticon project, and then point to examples how the concrete history of individual concept lists, in some sense their "evolution" can be studied. In doing so, I hope to emphasize the importance of a Concepticon resource (not necessarily our Concepticon) for studying how linguistic ideas evolve.

Types of Concept Lists

To get a better of a standing of the diversity of concept lists in the linguistic and scientific literature, it is useful to look at the various *types* of concept lists that have been produced in the past. We distinguish these types by using a specific tagging system, which currently distinguishes the following tags.

Tag	Description
acquisition	Concept lists related to studies on language acquisition.
annotated	Concept lists which contain further annotations which exceed the complexity of ranks.
areal	Concept lists designed for a specific linguistic area.
basic	Concept lists which are supposed to represent the basic vocabulary.
body parts	Concept lists which concentrate on body parts.
documentation	Concept lists which serve to document one language or one language family.
hihi	A list of highly reconstructable and highly retentive items (term from McMahon & McMahon 2005).
historical	A list which is historically interesting, mostly referring to lists published before the 20th century.
lolo	A list of less stable basic items, with low reconstructability and low retentiveness (term from McMahon & McMahon 2005).
naming test	A list designed for a naming test in neurology or psycholinguistics to asses the linguistic capability of children and adults.
proto-language	A list illustrating the concepts in a proto-language which can be reconstructed with high certainty.
questionnaire	A questionnaire for linguistic field work.
ranked	A list that shows items in a ranked order, and has one column reflecting the rank.
sign language	A list which was designed to investigate sign languages.
specific	A list that we deem specific, since it is not easy to compare with other lists in our sample.
stable	A list that is supposed to represent the stable part of a larger list. Usually, the stable part has an unstable counterpart.
ultra-stable	A usually very short list of the supposedly most stable concepts.
unstable	A list that is supposed to represent the unstable part of a larger list. Usually has a stable counterpart.

A given concept list can, of course, have more than a single tag, although some tags, such as “hihi” (highly stable sublists, following the framework described in McMahon et al. 2005) or “lolo” (low-stability sublists, as defined by ibid.), are usually not found to co-occur with additional tags.

Among our concept lists, those tagged as “specific” represent a dummy category for which we may create additional tags in the future. When filtering the currently available lists tagged in this form in the CLLD aplication, we can see that “body parts” are a good candidate for future concept lists that could be separated from the other “specific” lists.

Concept List	Compiler	Tags	Description	Source
Wilkins-1996-75	David P. Wilkins	specific, body parts	a collection of concepts, involving frequently recurring pathways of semantic shifts involving body parts	Wilkins 1996
Wilkins-1996-41	David P. Wilkins	specific, body parts	a collection of important body parts and liquids	Wilkins 1996
Snoek-2013-61	Conor Snoek	specific, body parts	a collection of body parts and liquids	Snoek 2013
Payne-1991-202	David L. Payne	specific	concept list underlying a collection of Proto-Arawakan lexical reconstructions	Payne 1991
Dolch-1936-220	E. W. Dolch	specific	“basic sight vocabulary, compiled for educational purposes	Dolch 1936
Goddard-2001-42	Cliff Goddard	specific	elicitation glosses for 42 semantic primes used in Natural Semantic Metalanguage	Goddard 2001
Dixon-1919-175	Roland B. Dixon and A. L. Kroeber	specific	collection of lexical cognates derived from a concept list of 225 English glosses not reported in the study	Dixon and Kroeber 1919
Mann-2004-118	Noel W. Mann	specific	a list of concepts that recur in other concept lists	

As we can see from the table, what is currently labelled as being “specific” could be easily further subdivided in the future. For example, the lists by Payne and Dixon and Kroeber represent endeavors in historical language comparison, during which the unity of original concept lists was broken, since the authors identified cross-linguistic cognates, but usually did not report their original questionnaire. If we manage to find more suitable concept lists illustrating Natural Semantic Metalanguage, we could add a tag for concept lists devoted to these studies, and if we add more concept lists designed for educational purposes, like the one proposed by Dolch, we could add a tag for “education”.

The largest group of tags in our collection (and this is not surprising, given our predominant interest in historical language comparison),are those tagged as “basic” or “stable”. The majority of these concept lists was published after Swadesh published his first concept lists, and their publication tradition is almost unbroken from the 1950s up to today. Given their importance for modern phylogenetic investigations in historical linguistics (Gray and Atkinson 2003, Chang et al. 2015), which — by default — make use of the same data that Swadesh envisioned originally (Kaplan 2017), it seems interesting to pay specific attention to their compilation history.

Evolution of Concept Lists

We can learn a lot about concept lists by simply reading the publications in which they were first announced. When paying specific attention to the elicitation glosses and the specific semantics they invoke, however, one will quickly realize that there is often a discrepancy between what scholars

name as the direct sources of their concept lists, and what they actually used. As a first example for these problems — resulting from a rather inaccurate treatment of semantics and elicitation glosses — we can look at the lists proposed by Swadesh himself.

In his first officially available publication from 1950, for example, Swadesh mentions a base list of 225 concepts, but in the very text where he lists all concepts, we find only as many as 215 elicitation glosses. The number 225 was — obviously — a typographical error. Unfortunately, this error was repeated by many scholars who did not count the concepts but rather took the paper by Swadesh at face value.

More interesting than the number of concepts first proposed by Swadesh is the fact that the concepts that Swadesh seeked to elicit themselves changed during the years in which Swadesh tried to further elaborate his theory of glottochronology, mostly unnoticed by Swadesh himself, or the people who tried to contradict or to support him. As one of the most interesting examples, consider the Concepticon concept set CHILD, which has two narrower concept sets, according to our underlying ontology, CHILD (DESCENDANT) and CHILD (YOUNG HUMAN). The rule for the links we make in our project is that we link to the more specific concepts in those cases where we are sure, but that we link to the broader concept, where the original concept list does not further specify a distinction between the two basic meanings of “child”.

When comparing the “reflexes”, i.e., the elicitation glosses in the different concept lists that were linked to either of the three concept sets, we can see that Swadesh’s elicitation glosses are linked to all three of them.

Concept List	Elicitation Gloss	Concept Set
Swadesh-1955-215	child	CHILD
Swadesh-1960-200	child	CHILD
Swadesh-1950-215	child (son or daughter)	CHILD (DESCENDANT)
Swadesh-1952-200	child (young person rather than as relationship term)	CHILD (YOUNG HUMAN)

Thus, Swadesh himself changed the concrete definition of what is often simply known as one of his basic concepts three times in his career, but surprisingly nobody really seems to have realized this, since later scholars would usually emphasize that both the Swadesh list of 100 items (from 1955) and the Swadesh list of 200 items (from 1952) would yield 207 concepts in total, while a thorough count in the form of Concepticon concept sets, where we refuse to identify cases like CHILD (YOUNG HUMAN) with CHILD (DESCENDANT) as reflecting the same concept, yields a union of 213 items, since six specifications differ enough to assign them to different concept sets, as shown in the table below.

Swadesh (1952)	Concepticon	Swadesh (1955)	Concepticon
skin (persons)	SKIN (HUMAN)	skin	SKIN
to rain	RAINING (RAINING)	rain	RAINING OR RAIN
man (male human)	MALE PERSON	man	MAN

Swadesh (1952)	**Concepticon**	**Swadesh (1955)**	**Concepticon**
warm (of weather)	WARM (OF WEATHER)	warm (hot)	HOT OR WARM
cold (of weather)	COLD (OF WEATHER)	cold	COLD
to burn (intrans)	BURNING	burn	BURN

While one could argue that Swadesh was more or less thinking of the same concepts and just specifying them differently, it is clear to me that the missing specifications in the list from 1955 do not provide any further evidence that would allow us to interpret elicitation glosses like “rain” narrowly as referring to the action (RAINING OR RAIN is unspecified as concept set in the distinction between “the rain” and “to rain”), or to interpret “burn” as inherently intransitive. This would specifically be misleading, since it was due to Swadesh’s unspecific elicitation glosses that the terms were later interpreted differently. Thus, in the concept lists by Starostin (1991 and later), Swadesh’s “burn” is consistently interpreted as a transitive term (BURN (SOMETHING) in the Concepticon), and “rain” is already specified as a noun (RAIN (PRECIPITATION)) in the Russian translation of Swadesh’s 1955 paper (published in 1964).

The discussion about the misinterpretation in Swadesh’s own work and by other scholars about the concrete values of the concepts used for lexicostatistic studies may seem extremely pedantic, especially when discussed in this detail. I find it nevertheless interesting and also important that we make clear to ourselves how lax people treated and still treat the role that comparative concepts play in lexical comparison. Given that lexicostatistic methods (including modern phylogenetic approaches) often receive harsh criticism from traditional scholars (Hoijer 1956, Pereltsvaig and Lewis 2015), it is interesting that traditional scholars themselves never pointed to the problems of concept mis-specifications that we can easily see when looking at the different versions of concept lists through the Concepticon resource.

Even more, it is fascinating to see how even renowned scholars commit classical beginner’s mistakes when trying to apply lexicostatistic techniques. An example are Winfred Lehman’s otherwise excellent “Exercises to accompany *Historical linguistics. An introduction*” (Lehmann 1962: 33), where he tries to illustrate how glottochronology works by translating Swadesh’s list of 100 items into four modern Indo-European languages, and then translates English *bark* as *bellen* (= “to bark”) in German and *écorce* (= “the bark”) in French (Lehmann’s list is not yet linked to Concepticon).

In a similar vein, we find scholar quoting Swadesh as their source, but using elicitation glosses that were obviously not taken from the three original publications by Swadesh. While this is not easy to prove in all cases (and we all know the problem of not having always the time to search for all original sources), we can find direct evidence for this when comparing specific peculiarities in terms of elicitation glosses used by different authors. One such peculiarity, first introduced by Gudschinsky’s (1956) influential summary of lexicostatistic techniques, is the use of a dash symbol in elicitation glosses such as “hold-take”, “fat-grease”, “right-correct”, “man-male”, “meat-flesh”, and “stab-pierce” (see the full list of Gudschinsky for details).

What is interesting about this peculiarity (since a dash symbol is otherwise rarely used to indicate alternative gloss-words) is that it was taken up in the again very influential survey on lexicostatistic methodology by Hymes (1960) that was accompanied by a long discussion about the lexicostatistic methodology involving many by then renowned experts. Hymes, however, reproduced Swadesh's different test lists with this practice as it was first introduced by Gudschinsky for the list shared by Hymes, and we find the exact same six concepts being glossed in the same way in Hyme's version of Swadesh's 200-item list from 1952 (see Swadesh-1960-200).

Searching in the Concepticon resource for lists that further copy this peculiarity, we find two more candidates (so far): the list by Gregersen (1976), which expands on the practice of "dashing" by adding elicitation glosses such as "spit-saliva" and "warm-hot", and the list by Nagaraja et al. (2013), which is to 98% identical with the list of Gudschinsky, as can be seen from the automatically calculated "similarity score" that we provide along with the CLLD application of the Concepticon, and also uses the dashes for all elicitation glosses originally introduced by Gudschinsky.

What is interesting in this context is that Gregersen quotes Hymes (1960), but says in the paper that he uses Swadesh's original concept lists, while the publication by Nagaraja et al. neither quotes Hymes, nor Gudschinsky, nor Swadesh himself, but just mentions the use of a "standard Swadesh list" to compare Mon-Khmer languages in the study. When re-reading the orignal publication by Nagaraja et al., I realized that the list of 200 concepts the authors use for their study itself goes back to an earlier publication by the first author from 2004, which was not accessible to me when writing this post, preventing me from verifying the origin of this 200-item list that looks so strikingly similar to Gudschinky's derivation of Swadesh's 200-item list.

It is difficult to tell what exactly happened, especially when trying to derive all this information from sources alone (without being able to actually ask the original concept list compilers). What these small and simply qualitative investigations presented in this context reveal, however, is that the seemingly simple "project" of compiling a concept list for the purpose of historical language comparison reveals an intertwined underlying history that may often be quite different from what we can read from the surface of the papers alone. Apart from a certain degree of carelessness in scholarly practice that these examples reveal, they also reflect the highly interesting dynamics underlying the scientific evolution of ideas. If we treat a concept list as something similar to a document that was copied and replicated through history, often with minimal errors introduced by different scholars, we can treat the study of the history of concept lists as a "stemmatic" enterprise, by which we try to trace the flow of information through time.

Summary and Conclusion

What, if anything, can we learn from the comparison of concept lists? In this post, I have introduced a project that tries to systematically link the various concept lists that have been compiled in the past. I have defended this project against different criticisms that were brought up in the past. But apart from the practical purpose of the Concepticon as a tool that helps us to aggregate datasets and increase the comparability of linguistic data, we can also use it as a resource that reflects how ideas are copied and spread from scholar to scholar. I have illustrated how this can be done by pointing to

some observations I made through qualitative data analysis in the past.

Given that the data is available in digital form, however, it should also be possible to investigate some of these questions automatically or within computer-assisted frameworks. Scholars may still ask themselves why I would pursue such an endeavor, and why it would be important to investigate how concept lists were shared and modified through the history of linguistics. Apart from general scientific curiosity, I see two aspects that are important in this regard. First, given that a resource like Concepticon can reveal problems in our linguistic practice of eliciting meanings and concepts, this endeavor can serve as a warning for future research, and as an appeal to scholars to be more explicit in our studies involving comparative concepts. Second, given that the "surface history" as reflected in citation practice by different scholars is obviously not always equivalent with the "underlying history" as revealed by a closer comparison of the data used by different scholars, it seems that a closer study of concept lists (but potentially also other concepts, including terminology, or comparative concepts for other kinds of linguistic data), may provide interesting insights into the sociological dynamics of our field.

References

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

Beaufils, V. (2015): eLinguistics.net. Quantifying the genetic proximity between languages.

Benedict, P. (1976): Sino-Tibetan: Another Look. *Journal of the American Oriental Society* 96.2. 167-197.

Bowern, C., P. Epps, R. Gray, J. Hill, K. Hunley, P. McConvell, and J. Zentz (2011): Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages?. *PLoS ONE* 6.9. e25195.

Bowern, C. and Q. Atkinson (2012): Computational phylogenetics of the internal structure of Pama-Nguyan. *Language* 88. 817-845.

Bowern, C. (2012): The riddle of Tasmanian languages. *Proceedings of the Royal Society of London B: Biological Sciences* 279.1747. 4590-4595.

Brinton, D. (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.

Brown, C., E. Holman, S. Wichmann, V. Velupillai, and M. Cysouw (2008): Automated classification of the world's languages. A description of the method and preliminary results. *Sprachtypologie und Universalienforschung* 61.4. 285-308.

Chang, W., C. Cathcart, D. Hall, and A. Garret (2015): Ancestry-constrained phylogenetic analysis ssupport the Indo-European steppe hypothesis. *Language* 91.1. 194-244.

Běijīng Dàxué 北京大学 (1964): Hànyǔ fāngyán cíhuì [Chinese dialect vocabularies]. Wénzì Gǎigé 文字改革:

Cross, E. (1964): Lexicostatistics has not yet attained the status of a science. In: Proceedings of the international congress of linguistics. 481-489.

Dixon, R. and A. Kroeber (1919): Linguistic families of California. University of California Press: Berkeley.

Dolch, E. (1936): A basic sight Vocabulary. *The Elementary School Journal* 36.6. 456-460.

Dolgopolsky, A. (1964): 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. 53-63.

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

Goddard, C. (2001): Lexico-semantic universals: A critical overview. *Linguistic Typology* 5. 1-65.

Goddard, C. (2010): The natural semantic metalanguage approach. In: Heine, B. and H. Narrog (eds.): The Oxford handbook of linguistic analysis. Oxford University Press: Oxford. 459-484.

Gray, R. and Q. Atkinson (2003): Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426.6965. 435-439.

Gregersen, E. (1976): The glottochronological performance of African languages. *Cahiers de l'Institut de Linguistique de Louvain* 3.5-6. 107-146.

Gudschinsky, S. (1956): The ABC's of lexicostatistics (glottochronology). *Word* 12.2. 175-210.
Haspelmath, M. (2010): Comparative concepts and descriptive categories. *Language* 86.3. 663-687.
Hoijer, H. (1956): Lexicostatistics. A critique. *Language* 32.1. 49-60.
Hymes, D. (1960): Lexicostatistics So Far. *Current Anthropology* 1.1. 3-44.
Kaplan, J. (2017): From lexicostatistics to lexomics: Basic vocabulary and the study of language prehistory. *OSIRIS* 32.1. 202-223.
Kiss, G., C. Armstrong, R. Milroy, and J. Piper (1973): An associative thesaurus of English and its computer analysis. In: Aitken, A., R. Bailey, and N. Hamilton-Smith (eds.): The computer and literary studies. Edinburgh University Press: Edinburgh. 153-165.
Lehmann, W. (1962): Exercises to accompany ''Historical linguistics. An introduction''. Holt, Rinehart and Winston: New York.
von Leibniz, G. (1768): Desiderata circa linguas populorum, ad Dn. Podesta [Desiderata regarding the languages of the world]. In: Dutens, L. (ed.): 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].6.2. Fratres des Tournes: Geneva. 228-231.
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. 2393-2400.
List, J.-M., S. Greenhill, C. Anderson, T. Mayer, T. Tresoldi, and R. Forkel (2018): CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats. *Linguistic Typology* 22.2. 277-306.
McMahon, A. and R. McMahon (2005): Language classification by numbers. Oxford University Press: Oxford.
McMahon, A., P. Heggarty, R. McMahon, and N. Slaska (2005): Swadesh sublists and the benefits of borrowing: An Andean case study. *Transactions of the Philological Society* 103. 147-170.
Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016): A Central Asian language survey. *Language Dynamics and Change* 6.1. 57–98.
Nagaraja, K., P. Sidwell, and S. Greenhill (2013): A lexicostatistical study of the Khasian language. *Mon-Khmer Studies* 42. 1-11.
Nicholas, L., R. Brookshire, D. MacLennan, J. Schumacher, and S. Porrazzo (1989): The Boston Naming Test: Revised administration and scoring procedures and normative information for non-brain-damaged adults. In: Clinical Aphasiology Conference. 103-115.
Ogden, C. (1930): Basic English: A general introduction with rules and grammar. Kegan Paul: London.
Pallas, P. (1786): Modèle du Vocabulaire, qui doit servir à la comparaison de toutes les langues A vocabulary model which may serve the comparison of all languages. Sankt Pétersbourg.
Pallas, P. (1789): Sravnitel'nye slovari vsech jazykov i narečij, sobrannye desniceju Vsevysočajšeij Osoby. Otdelenie pervoe, soderžaščee v sebe evropejskie i aziatskie jazyki [Comparative dictionaries of all languages and all speeches. Collected under supervision of the Queen. Part one, containing European and Asian languages.]. Šnor: Saint Petersburg.
Payne, D. (1991): A classification of Maipuran (Arawakan) languages based on shared lexical retentions. In: Derbyshire, D. and G. Pullum (eds.): Handbook of Amazonian languages.3. Mouton de Gruyter: Berlin and New York. 355-499.
Pereltsvaig, A. and M. Lewis (2015): The Indo-European Controversy. Facts and fallacies in historical linguistics. Cambridge University Press: Cambridge.
Poornima, S. and J. Good (2010): Modeling and encoding traditional wordlists for machine applications. In: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground.. Association for Computational Linguistics 1-9.
Shevoroshkin, V. and A. Manaster Ramer (1991): Some recent work on the remote relations of languages. In: Lamb, S. and E. Mitchell (eds.): . Standorf University Press: Stanford. 178–199.
Snoek, C. (2013): Using semantically restricted word-lists to investigate relationships among Athapaskan languages. In: Borin, L. and A. Saxena (eds.): Approaches to Measuring Linguistic Differences. Mouton de Gruyter: Berlin. 231-248.
Starostin, S. (1991): Altajskaja problema i proischo\vzdenije japonskogo jazyka [The Altaic problem and the origin of the Japanese language]. Nauka: Moscow.
Société de Linguistique de Paris (1871): Statuts. Approuvés par décision ministérielle du 8 Mars 1866. *Bulletin de la Société de Linguistique de Paris* 1. III-IV.
Swadesh, M. (1950): Salish internal relationships. *International Journal of American Linguistics* 16.4. 157-167.
Swadesh, M. (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.
Swadesh, M. (1955): Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*

21.2. 121-137.
Swadesh, M. (1964): K voprosy o povyshenii to\ccnosti v leksikostatisti\cceskom datirovanii. *Novoe v Lingvistike* 1. 53-87.
Wilkins, D. (1996): Natural tendencies of semantic change and the search for cognates. In: Durie, M. (ed.): The comparative method reviewedThe comparative method reviewed. Regularity and irregularity in language change. Oxford University Press: New York. 264-304.
Haspelmath, M. and U. Tadmor (eds.) (2009): World Loanword Database. Max Planck Digital Library: Munich. http://wold.livingsources.org.
University, P. (2010): WordNet. A lexical database for English. Online Resource.
Youn, H., L. Sutton, E. Smith, C. Moore, J. Wilkins, I. Maddieson, W. Croft, and T. Bhattacharya (2016): On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences.*

How to cite this post:

List, Johann-Mattis. 2018. Towards a history of concept list compilation in historical linguistics. *History and Philosophy of the Language Sciences*. http://hiphilangsci.net/2018/10/31/concept-list-compilation/