Academic Journal of Modern Philology

Matthias Urban
University of Tübingen

ISSN 2299-7164
Vol. 7 (2018)
s. 141–158

Towards a Semantically Organized Meaning List for Cognate Searches

Abstract

In this contribution I present a meaning list for cognate searches in which meanings are, unlike in similar lists of "basic vocabulary" for the same purpose, organized according to semantic principles. This list is designed to identify possible cognates, which can then be scrutinized more closely in search for hitherto undetected genealogical relationships between languages, in a more effective way. Rather than proposing a completely new set of meanings to be featured, the list combines those most commonly used in extant lists of basic vocabulary such as the Swadesh or the Leipzig/Jakarta List, but introduces several design principles which are jointly able to represent also complex semantic relationships in the context of wordlists.

Keywords: meaning list, meaning, semantics, semantic relationships.

1. Rationale

Of all the steps the historical linguist needs to take to establish hitherto undemonstrated genealogical relations between languages or language families, the initial one, in which items that may be comparable are assembled, is the least subject to methodical rigor. Indeed, the comparative method of historical linguistics is a powerful tool to *test* hypothesis about genealogical relationships, but it is unsuitable to *generate* them (Weiss 2014: 128). Accordingly, textbook demonstrations of the comparative method such as that by Campbell (2013: 107–128) usually begin with ready-made sets of (putative) cognates to which the method is then applied, largely ignoring the question as to how one arrives at such ordered sets.

How, then, are hypotheses regarding language relationships generated? As Dolgopolsky (1986: 27) puts it, one carries out a "preliminary assessment of the advisability of making the comparison" between a set of languages and/or families. This involves the examination of "basic core lexemes and grammatical morphemes, but initially only at a superficial, non-etymological level." Perhaps more likely than this "out of the blue" way to proceed, however, is that a researcher already harbors some idea regarding a possible relationship between two specific languages or language families, for instance, because s/he

has casually observed lexical similarities. Another way now available is to employ quantitative statistical methods as heuristics for possible genealogical relationships (Wichmann et al. 2010). Again, however, for a relationship to be actually demonstrated, lexical items will need to be sifted manually again in search for possible cognates which can then be subjected to the methodological rigor of the comparative method. Meanwhile, computational techniques for automated cognate detection are advancing rapidly (e.g. Hall and Klein 2010, Steiner et al. 2011, List 2014, Jäger and Sofroniev 2016), but they are as of yet likely still vulnerable to those processes of language change that obscure cognate relationships in the first place, viz. fossilized morphology, historical processes of word-formation (cf. the difficulties with Chinese data reported in List et al. 2017), and, indeed, semantic change. For the time being, therefore, human judgment remains indispensable. Just like recent computational algorithms seek to assist humans in cognacy assessments rather than replace human judgement (List et al. 2017), here I present a non-computational aid for cognate searches which pays particular attention to factoring in semantic change.

2. Wordlists in cognate searches

The lexica of languages are vast, to the effect that typically researchers restrict their attention to a particular set of lexical items for initial comparison, the so-called "basic vocabulary". This is the part of the lexicon which is assumed to be relatively time-stable and more likely to be inherited than replaced either internally or borrowed. Here, wordlists of such "basic vocabulary" come into play. Such wordlists, of course, have their problems, some related exclusively to their use in the contentious technique of glottochronology, some of a more general nature (e.g. McMahon and McMahon 2005: 40–44, Campbell 2013: 453–456). I shall not be concerned further with these issues here; instead I would like to focus on two weaknesses of wordlist-based comparison that are virulent in particular if one is employing them in attempts to establish initial evidence for genealogical relationships.

Items on wordlists are ordered by glosses, with items having the "same meaning" –a considerable simplification in most cases– appearing in the same row. These are then compared in search of "lookalikes" (even though the comparative method actually does not require that valid comparanda in fact look alike) to be checked later for regular correspondences, or other evidence that seems worthwhile to subject to more rigorous comparison later. Divergence from a common source for which evidence is sought, however, happens along two major axes, corresponding to the two major sides of linguistic signs: one is the formal dimension, i.e. the shape of words, which is subject to sound change and phonological restructuring. Even though we are not yet in possession of an empirical catalogue of attested sound changes against which to assess the plausibility of comparisons (though see Kümmel 2007 for something close to this for consonants), historical linguists typically are able to tell which sound changes are "natural" and, if regular, constitute plausible evidence for divergence from a common ancestor. The other dimension is the semantic one. This terrain is much less secure, even though there has been work in the area (see Urban 2014 for an overview). However, just like sound change, semantic change is an undeniable reality of language change and must hence be reckoned with in the search for possible cognates.

Furthermore, just like current computational approaches, the manual use of wordlists is prone to missing "hidden" cognates (Koch and Hercus 2013), i.e. the situation when cognate relationships are superficially obscured by occurring in one or more of the languages only as fossilized forms, bases of

derivatives, members of compounds, etc. which also make automated cognate detection still a difficult enterprise. To the degree that the purported relationship becomes less shallow, I imagine the number of cognates only available in this "hidden" form to grow accordingly.

A semantic reorganization of wordlists has the potential to help in both cases: on the one hand, words related semantically will inevitably appear immediately or at least closely together on such a reorganized list, thus facilitating their visual detection by human inspectors. On the other hand, "hidden" cognates, at least in a salient subclass of cases, will also appear in semantically related items (though see a cautionary note in section 4) and thus with a certain likelihood adjacent to one another. The following section describes the basic properties of such a semantically organized list which I currently use for research on the genealogical relationships of South American languages.

3. A semantically organized wordlist

3.1. Background

The wordlist I present in this article is in fact a rather humble contribution in that it merely attempts to put existing theoretical knowledge into practice rather than to establish new knowledge. Neither do I propose an entirely new list which includes different lexical items from those found on other lists. Nor is the very idea of semantically organized wordlist new: this credit goes to Wilkins (1996). In this seminal paper, Wilkins investigates regularities of semantic change in the domain of body-part terms across several language families. His data allow for several generalizations. For instance, based on inspection of several etymological dictionaries for different language families, Wilkins (1996: 276) establishes synecdochic continua in diachrony such as 'navel'→'belly'→'trunk'→'body'→'person', wherein terms denoting meanings to the left of each arrow may come to denote those to the right, but not the other way around. Wilkins suggests that meanings appearing next to each other in such chains–because they are demonstrably related by semantic change in more than one language family–could be ordered in wordlists for cognate searches in such a way that they appear next to each other there, too. Wilkins's data, however, are by design exclusively applicable to body-part terms (plus some non-body part terms which turn out to frequently be associated with them by semantic change). Body-part terms are without a doubt an important part of "core" vocabulary, but not the only one. It is therefore desirable to expand the idea of semantically organizing wordlists to be able to arrange an entire set of "core" vocabulary items along similar principles. This is the goal of the wordlist presented here.

3.2. The meanings covered

I use a combination of the Swadesh 100 and Swadesh 200 list (as represented in Campbell 2013: 449–451), pl us the recent Leipzig/Jakarta list of 100 items (Tadmor et al. 2010) to generate sets of meanings. The latter list is the result of a project which aimed to assess the variable borrowability of meanings in a variety of the world's languages; the 100 meanings are those for which words were least frequently borrowed, least frequently analyzable, but most widely represented and oldest in the languages investigated, thus having the desirable diachronic stability and (near-)universality expected for cognates. Contrary to what

one might assume, the Swadesh 100 list is not exactly a subset of the Swadesh 200 list, but contains some meanings not covered in the latter list. Thus, combining the two results in a number of meanings a little larger than 200. On the other hand, there is a large overlap between meanings on the Swadesh lists and the Leipzig/Jakarta list, but also here some new meanings need to be added to a combined list (cf. Tadmor et al. 2010: 242, table 8 for a list of the non-overlapping items). All in all, the combined list features 225 unique meanings.

Within computational linguistics and statistically-minded historical linguistics, there has been a tendency to further reduce the size of wordlists, for fear that either additional material beyond a presumed highly stable set of "core" items waters down the phylogenetic signal or at least does not contribute anything to sharpen it (see Heggarty 2010 for review). For traditional work, I believe that 200 items are a minimum for initial comparisons, and if anything the list should be longer rather than shorter. This is particularly so because hypotheses emerging from word-list comparisons must be confirmed and refined by means of the comparative method against a much larger set of data at a later stage of research anyway (Kaufman 1990: 18 considers 500-600 items of basic vocabulary and "[a]bout 100 points of grammar" necessary).

I should like to point out that I do not wish to make any claim as to a superior usefulness of precisely those meanings on the list. The combination of the Swadesh lists and the Leipzig/Jakarta list was born out of personal preferences in practical work in South America and an intuition that for exploratory manual work, an expanded rather than a reduced list is more suitable; I have no empirical data to back this up, nor do I want to convince anyone to follow my example in using a combined Swadesh/Leipzig-Jakarta list. My point lies in the restructuring of this particular list (or any other list other historical linguists may prefer in their work) according to semantic principles.

3.3. Sources of information on semantic connections

I use the following sources to restructure the organization of the meanings semantically:

(i) information on synchronic semantic associations in a broad sample of the languages of the world provided by the CLICS database of cross-linguistic colexifications (List et al. 2014). Colexification refers to the situation of two, ideally semantically related, meanings being expressed by the same item.[1] For present purposes, cases have been counted if they recur in more than one of the world's language families to reduce the possibility of accidental homonymy. Since synchronic polysemy is an intermediate stage in a salient subset of semantic change processes, colexification—which terminologically includes polysemy—is a usable proxy for semantic change (though see section 4 for some caveats). CLICS data were manually checked for unexpected patterns and cases due to conversion errors removed from consideration. Furthermore, it is vital to point out that some associations from CLICS have been—subjectively—not taken into account. This is for instance the case for the connection between 'leaf' and 'year'. While this may be a true semantic association—after all, it recurs in two independent language families covered in the CLICS database—it is thus rather rare and possibly spurious. The discarding of such information on subjective

1 Colexification as a technical term was coined by François (2008). It is something entirely different from the similar term "co-lexicalization" used in the work of Givón (e.g. 2009).

grounds is certainly debatable, but I have chosen to do so nevertheless in order to reduce the complexity of the data that need to be represented. Discarded connections are noted in the appendix after the complete meaning list itself.

(ii) information from Wilkins (1996: 284, table 10-1) concerning body-part terms. Wilkins's own arrangement of his data is ingenious in that chains of semantic development are brought to light to the effect that one might somehow measure semantic divergence according to how far meanings are removed on a common trajectory (cf. Wilkins 1996: 297). In my adaptation of his data, I have only taken into account direct links to avoid a further layer of complexity.

(iii) the directional pathways in semantic change suggested by Urban (2011) and,

(iv) to a very limited extent, generalizations from grammaticalization theory (such as the grammaticalization path DEMONSTRATIVE > 3rd PERSON PRONOUN, Heine and Kuteva 2002: 112–113).

The sources all contribute unique data to be integrated, but reassuringly, there is also a considerable overlap and mutual support for the associations they indicate.

3.4. Organizational principles of the list

The basic organizational principle of the list, which can be found in the appendix to this paper, is the adjacency principle: this simply says that meanings appearing immediately on top and below of each other are to be interpreted as semantically related according to the above sources. Respective terms in the languages one wishes to investigate should accordingly be checked for possible cognacy just like semantically isomorphic items.

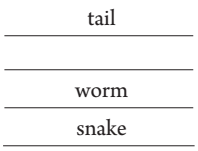

Fig. 1: a first excerpt from the semantically reorganized list.

Thus, in the excerpt from the list in (1), the adjacency principle says that words for 'worm' and 'snake' should be checked for properties that may point to cognacy, whereas this is not the case for 'tail', which is separated from the other meanings by a blank line. Actually, 'tail' does not merely happen to appear in the same area: the list was designed in such a way that items or clusters of items for which a semantic commonality can be perceived–e.g. animal and body part-terms–appear roughly in the same area of the list even when the sources used for semantic organization do not indicate such a relationship. This is merely a measure to avoid a chaotic feel of the list, and does not mean that all items in such broader domains should be compared to one another.

It is frequently the case that semantic associations are not found for two single meanings as in the above case, but that larger clusters of related meanings emerge, sometimes with considerable internal complexity. Consider by way of example the community network from the CLICS database in (2).

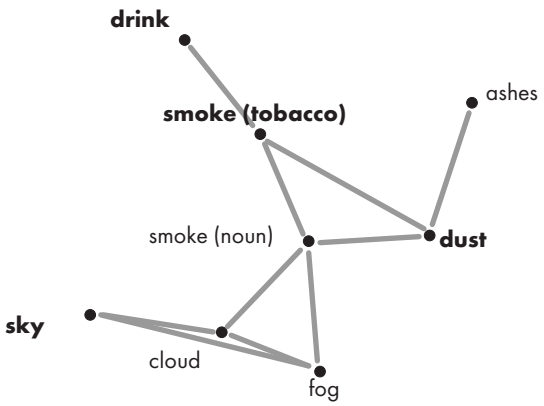

Fig. 2: a community network from CLICS.

This community network, which is already simplified according to CLICS's procedures (List et al. 2013) and which actually forms part of a much larger network, shows multiple connections for some of the meanings. 'Smoke', for instance, is related to 'cloud', 'fog', and 'dust'. The mere principle of adjacency is unable to handle even this moderately complex case. If the situation were to be translated into the organization of a wordlist one would need a multidimensional list in which related meanings can branch off into several directions, more than two anyway. Yet wordlists are two-dimensional, and the principle of adjacency only permits two places –above and below a specific item– where related items can "dock". From the network one can see that the meanings 'cloud', 'fog', and 'smoke' turn out to be all semantically related to one another. Hence, a representation as in (3) would not represent all information: it tells us to compare terms for 'cloud' with those for 'fog', and those for 'fog' in turn with those for 'smoke', but it does not tell us to also compare 'cloud'-terms with 'smoke'-terms.

cloud
fog
smoke (n.)

Fig. 3: a second excerpt from the semantically reorganized list

This calls for an additional way of representation, which is a curly bracket to the left of a set of meanings, as in (4). The curly bracket indicates that terms for all meanings within the set should be compared with all others.

cloud
fog
smoke (n.)

Fig. 4: a third excerpt from the semantically reorganized list, illustrating the use of curly brackets.

Yet matters are even more complicated than suggested by the network in (4): it does not show that 'dust', which appears to be rather peripheral in the network, is, alongside 'ashes', also connected to 'earth (soil)' and 'sand'. In fact, 'dust', 'earth (soil)' and 'sand' form the same kind of "triplet" as 'cloud', 'fog', and 'smoke'. Yet 'dust' is also connected to the rather isolated 'ashes', but the other meanings are not. If one wants to represent this latter association of 'ashes' and 'dust' at the same time as the other information, one is forced to break up the triplet, and the information becomes misrepresented. To handle such cases, another representation technique, indentation, is introduced. Meanings that appear indented, even when occurring in triplets, are to be interpreted as being connected only with the meaning immediately above and to be ignored otherwise. In a few cases, more than one meaning appears indented adjacent to each other. This does not change the rules; these intended meanings do not need to be compared to one another. An example of a cluster featuring an indented meaning, to be read as "compare items meaning 'ashes' with those for 'earth=ground, soil', and compare items meaning 'dust', 'ashes', and 'sand' with one another" is in figure (5).

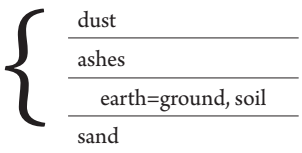

Fig. 5: a fourth excerpt from the semantically reorganized list, illustrating the use of indentation.

The full cluster as it actually appears in the wordlist looks as in (6).

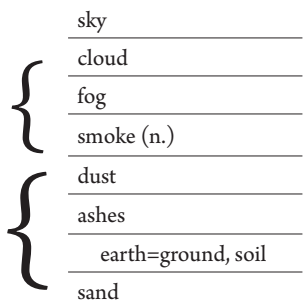

Fig. 6: a fifth excerpt from the semantically reorganized list, illustrating a complex cluster of related meanings on the list.

Finally, some meanings are repeated in more than one place on the list, as is also the case in Wilkins's (1996: 284, table 10-1) representation of his more limited data. Repetition becomes useful if meanings take part of two recognizable and distinguishable clusters of semantic relatives which already in themselves are rather complex.

4. Outlook

Obviously, the approach used here to arrive at a semantically organized wordlist for cognate searches leaves much to be wanted and wide room for further improvement and refinement. I would like to specifically mention three cases in point: first and foremost, even though there is an empirical grounding of the ordering, there are nevertheless few cases where I have subjectively considered a particular association as possibly spurious and hence ignored. Further objectification would be desirable. Second, the repetition of some elements in various places of the list is a workable, but probably not yet the ideal solution. Third, while colexification is a reasonable proxy to semantic change, it must be pointed out that for cognate search more generally, exclusive reliance on colexification may sometimes be insufficient. For instance, it is not uncommon for languages of the world to express antonyms using partly the same morphological material, one of the antonyms being expressed as a negation of the other (e.g. 'narrow' = 'not wide' etc.). Under the hypothetical situation that the form expressing 'wide' is replaced through time in a language, the cognate survives in "hidden" form only as part of its antonym. Yet in CLICS, there is, as one may have guessed, no language which colexifies 'narrow' and 'wide', for which reason the two meanings are not associated on the present list. On the long run, it would therefore be beneficial if the organization of the wordlist would be amended to take factors such as this into account.

Nevertheless, I believe that even in its present form the list may benefit exploratory searches for cognates. Computational approaches are by no means ignorant of the reality of semantic change; Kondrak (2009) uses WordNet, Steiner et al. (2011) an approach inspired by semantic maps to take it into account. Perhaps the information contained in the present list–or, more broadly, its sources–can in the future be incorporated in automated procedures to further improve them.

References

Campbell, Lyle (2013) *Historical linguistics. An introduction*. 3rd ed. Edinburgh: Edinburgh University Press.

Dolgopolsky, Aaron B. (1986) "A probabilistic hypothesis concerning the oldest relationships among the language families in Northern Eurasia." [In:] Vitalij V. Shevoroshkin, Thomas L. Markey (eds. and transl.) *Typology, relationship, and time. A collection of papers on language change and relationship by Soviet linguists*. Ann Arbor: Karoma; 27-50.

François, Alexandre (2008) "Semantic maps and the typology of colexification: intertwining polysemous networks across languages." [In:] Martine Vanhove (ed.) *From polysemy to semantic change*. Amsterdam/Philadelphia: John Benjamins; 163-215.

Givón, T. (2009) *The genesis of syntactic complexity: diachrony, ontogeny, neuro-cognition, evolution*. Amsterdam/Philadelphia: Benjamins.

Hall, David, Dan Klein (2010) "Finding cognate groups using phylogenies." [In:] *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics; 1030-1039.

Heggarty, Paul (2010) "Beyond lexicostatistics: how to get more out of 'word list' comparisons." *Diachronica* 27(2); 301-324.

Heine, Bernd, and Tania Kuteva (2002) *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.

Jäger, Gerhard, Pavel Sofroniev (2016) "Automatic cognate classification with a support vector machine." [In:] Stefanie Dipper, Friedrich Neubarth, Heike Zinsmeister (eds.) *Proceedings of the 13th Conference on Natural Language Processing (KONVENS), Bochum, Germany, September 19-21*. Bochum: Sprachwissenschaftliches Institut, Ruhr-Universität Bochum; 128-134

Kaufman, Terrence (1990) "Language history in South America: what we know and how to know more." [In:] Doris L. Payne (ed.) *Amazonian linguistics. Studies in lowland South American languages*. Austin: University of Texas Press; 13-73.

Koch, Harold, Luise Hercus (2013) "Obscure vs. Transparent cognates in linguistic reconstruction." [In:] Robert Mailhammer (ed.) *Lexical and structural etymology: beyond word histories*. Boston/Berlin: Walter de Gruyter; 33-52.

Kondrak, Grzegorz (2009) "Identification of cognates and recurrent sound correspondences in word lists." *Traitement automatique des langues* 50 (2); 201-235.

Kümmel, Martin (2007) *Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels und ihre Konsequenzen für die vergleichende Rekonstruktion*. Wiesbaden: Reichert.

List, Johann Mattis (2014) *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis, Anselm Terhalle, Matthias Urban. 2013. "Using network approaches to enhance the analysis of cross-linguistic polysemies." [In:] *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Potsdam: Association for Computational Linguistics; 347-353.

List, Johann-Mattis, Thomas Mayer, Anselm Terhalle, Matthias Urban (2014) *CLICS: Database of Cross-Linguistic Colexifications*. Version 1.0. Marburg: Forschungszentrum Deutscher Sprachatlas. Online: http://CLICS.lingpy.org.

List, Johann-Mattis, Simon J. Greenhill, Russell D. Gray (2017) "The potential of automatic word comparison for historical linguistics." *PLoS ONE* 12(1): e0170046. https://doi.org/10.1371/journal.pone.0170046

McMahon, April, Robert McMahon (2005) *Language classification by numbers*. Oxford: Oxford University Press.

Steiner, Lydia, Peter F. Stadler, Michael Cysouw (2011) "A pipeline for computational historical linguistics." *Language Dynamics and Change* 1(1); 89-127.

Tadmor, Uri, Martin Haspelmath, Bradley Taylor (2010) "Borrowability and the notion of basic vocabulary." *Diachronica* 27 (2); 226-246.

Urban, Matthias (2011) "Asymmetries in Overt Marking and Directionality in Semantic Change." *Journal of Historical Linguistics* 1; 3-47.

Urban, Matthias (2014) "Lexical semantic change and semantic reconstruction." [In:] Claire Bowern, Bethwyn Evans (eds.) *The Routledge Handbook of Historical Linguistics*. Abingdon/New York: Routledge; 374-392.

Weiss, Michael (2014) "The comparative method." [In:] Claire Bowern, Bethwyn Evans (eds.) *The Routledge Handbook of Historical Linguistics*, Abingdon/New York: Routledge; 127-145.

Wichmann, Søren, Eric W. Holman, André Müller, Viveka Velupillai, Johann-Mattis List, Oleg Belyaev, Matthias Urban, Dik Bakker (2010) "Glottochronology as a heuristic for genealogical language relationships." *Journal of Quantitative Linguistics* 17 (4); 303-316.

Wilkins, David P. (1996) "Natural tendencies of semantic change and the search for cognates."[In:] Mark Durie, Malcolm Ross (eds.): *The comparative method reviewed: regularity and irregularity in language change*. Oxford: Oxford University Press; 264-304.

Appendix: A semantically aligned list with cross-references to the Leipzig/Jakarta list (Tadmor et al. 2010: 239–241, table 7) and the Swadesh-100 and Swadesh-200 lists as represented in Campbell (2013: 449–451)

Semantically aligned list		**Leipzig/Jakarta**		**Swadesh-100**		**Swadesh-200**	
1	I	14	1SG pronoun	1	I	78	I
2	we			3	we	182	we
3	you (singular)	9	2SG pronoun	2	you	168	thou/yousingular
4	you (plural)					198	ye
5	they					163	they
6	he/she/it	34	3SG pronoun			67	he
7	this	38	this	4	this	167	this
8	that			5	that	161	that
9	one	32	one	11	one	109	one
10	few					46	few
11	some					146	some
12	two			12	two	176	two
13	three					169	three
14	four					57	four
15	five					50	five
16	hand	19	arm/hand	48	hand	66	hand
17	claw			45	claw		
18	every, all			9	all (of a number)	1	all
19	person, human being			18	person	111	person
20	man			17	man	94	man (male)
21	husband					77	husband
22	father					43	father
(19)	person, human being			18	person	111	person
23	wife					190	wife
24	mother					97	mother
25	child	51	child (kinterm)			20	child (young)
26	woman			16	woman	195	woman
27	dog	84	dog	21	dog	30	dog
28	tail	84	tail	35	tail	160	tail

	Semantically aligned list	Leipzig/Jakarta		Swadesh-100		Swadesh-200	
29	worm					197	worm
30	snake					144	snake
31	fly	20	fly				
32	ant	71	ant				
33	louse	15	louse	22	louse	93	louse
34	fish (n.)	38	fish	19	fish (noun)	49	fish
35	meat	18	flesh/meat	29	flesh (meat)		
36	animal					3	animal
37	bird	91	bird	20	bird	12	bird
38	wing	17	wing			192	wing
39	feather			36	feather	45	feather (large)
40	body hair[b]	31	hair	37	hair	65	hair
41	leaf	64	leaf	25	leaf	86	leaf
42	grass					62	grass
43	root	9	root	26	root	121	root
44	woods, forest					196	woods
45	tree			23	tree	174	tree
46	wood[a]	80	wood				
47	staff, walking stick					153	stick (of wood)
48	grease, fat			32	grease (fat)	42	fat (substance)
49	liver	66	liver	53	liver	91	liver
50	inside, in	97	in			81	in
51	heart			52	heart	70	heart
52	breathe, breath					18	tobreathe
53	suck	67	tosuck			156	tosuck
(50)	inside, in	97	in			81	in
54	stomach			49	belly	10	belly
(50)	inside, in	97	in			81	in
55	intestines, guts					64	guts
56	navel	42	navel				

	Semantically aligned list	**Leipzig/Jakarta**		**Swadesh-100**		**Swadesh-200**	
57	neck	23	neck	50	neck	103	neck
58	mouth	5	mouth	42	mouth	99	mouth
59	tooth	28	tooth	43	tooth	173	tooth (front)
60	tongue	6	tongue	44	tongue	172	tongue
61	bone	7	blood	31	bone	17	bone
62	horn	38	horn	34	horn		
63	leg	37	leg/foot			88	leg
64	walk			65	walk	178	towalk
65	thigh	76	thigh				
66	foot	37	leg/foot	46	foot	56	foot
67	hide, conceal	67	tohide				
68	skin, hide	67	skin/hide	28	skin	137	skin (of person)
69	bark			27	bark (of a tree)	8	bark (of a tree)
70	back	46	back			6	back
71	hard	99	hard				
72	knee	59	knee	47	knee		
(52)	breathe, breath					18	to breathe
73	blow	79	toblow			16	to blow (wind)
74	strike (hit, beat)	36	tohit/beat			73	tohit
75	wind	48	wind			191	wind (breeze)
76	sky					138	sky
77	year					199	year
78	day					26	day (not night)
79	earth =ground, soil	63	soil	79	earth (soil)	36	earth (soil)
80	sun			72	sun	157	sun
81	name	15	name	##	name	100	name
82	moon			73	moon	96	moon
83	yesterday	41	yesterday				
84	star	97	star	74	star	152	star
85	freeze					58	tofreeze

	Semantically aligned list	**Leipzig/Jakarta**		**Swadesh-100**		**Swadesh-200**	
86	night	20	night	92	night	105	night
87	shade, shadow	91	shade/shadow				
88	cold			94	cold	22	cold (weather)
89	ice					79	ice
90	snow (n.)					145	snow
91	fog					55	fog
92	salt	91	salt			125	salt
93	sea					129	sea (ocean)
94	lake					84	lake
(50)	inside, in	97	in			81	in
95	river, stream, brook					119	river
(53)	suck	67	to suck			156	to suck
96	drink	42	to drink	54	drink	31	to drink
97	water	4	water	75	water	181	water
98	rain (n.)	13	rain	76	rain	115	to rain
99	cloud			80	cloud	21	cloud
(77)	year					199	year
(76)	sky						
99	cloud			80	cloud	21	cloud
(91)	fog					55	fog
100	smoke (n.)	49	smoke	81	smoke	142	smoke
101	dust					34	dust
102	ashes	84	ash	83	ash(es)	4	ashes
(79)	earth=ground, soil	63	soil	79	earth (soil)	36	earth (soil)
103	sand	59	sand	78	sand	126	sand
104	stone, rock	27	stone/rock	77	stone	154	stone
105	mountain, hill			86	mountain	98	mountain
(62)	horn	38	horn	34	horn		
106	head			38	head	68	head
107	ear	22	ear	39	ear	35	ear
108	hear	61	tohear	58	hear	69	tohear
109	eye	83	eye	40	eye	39	eye
110	see	89	tosee	57	see	130	tosee
111	fruit					59	fruit
112	flower					53	flower
113	seed			24	seed	131	seed
114	egg	52	egg	33	egg	38	egg

	Semantically aligned list	**Leipzig/Jakarta**		**Swadesh-100**		**Swadesh-200**	
115	house	26	house				
(52)	breathe, breath					18	to breathe
116	smell (v. trans.)					141	to smell (perceive odour)
117	nose	2	nose	41	nose	106	nose
(108)	hear	61	to hear	58	hear	69	to hear
118	know	58	to know	59	know	83	know (facts)
(110)	see	89	to see	57	see	130	to see
119	do, make	25	to do/make				
120	give	53	to give	70	give	60	to give
121	say	28	to say	71	say	127	to say
122	think (= reflect)					166	to think
123	count					24	to count
124	laugh	61	to laugh			85	to laugh
125	play					112	to play
126	sing					135	to sing
127	cry, weep	87	to cry/weep				
128	fear, fright					44	to fear
129	sleep			60	sleep	139	to sleep
130	lie down			67	lie (down)	89	to lie (on side)
131	live, living, life					90	to live
132	sit			68	sit	136	to sit
133	stand	45	to stand	69	stand	151	to stand
134	dig					28	to dig
135	scratch					128	scratch (itch)
136	rub, wipe					124/193	rub/wipe
137	wash					180	to wash
138	breast (of woman)	12	breast	51	breast (female)		

	Semantically aligned list	Leipzig/Jakarta		Swadesh-100		Swadesh-200	
(53)	suck	67	to suck			156	to suck
(52)	breathe, breath					18	to breathe
139	pull					113	to pull
140	push, shove					114	to push
141	crush, grind	100	to crush/grind				
142	split					148	to split
143	cut					25	to cut
(74)	strike (hit, beat)	36	to hit/beat			73	to hit
144	stab					150	to stab (stick)
(141)	crush, grind	100	to crush/grind				
145	kill			62	kill	82	to kill
146	die, dead			61	die	27	to die
147	hunt					76	to hunt (game)
(119)	do, make	25	to do/make				
148	fight (v.)					47	to fight
(74)	strike (hit, beat)	36	to hit/beat			73	to hit
(119)	do, make	25	to do/make				
(144)	stab					150	to stab (stick)
(141)	crush, grind	100	to crush/grind				
149	squeeze, wring					149	to squeeze
(53)	suck						
(96)	drink	42	to drink	54	drink	31	to drink
(74)	strike (hit, beat)	36	to hit/beat			73	to hit
150	take	71	to take				
151	eat	75	to eat	55	eat	37	to eat
152	bite	46	to bite	56	bite	13	to bite
153	burn (v. intrans.)	53	to burn (intr.)	84	burn	19	to burn (intransitive)
154	hot			93	hot		
155	warm					179	warm (weather)
156	fire	1	fire	82	fire	48	fore
157	red	64	red	87	red	116	red
158	blood	7	blood	30	blood	15	blood
159	yellow			89	yellow	200	yellow
160	green			88	green	63	green
161	white			90	white	187	white
(117)	nose	2	nose	41	nose	106	nose
162	throw					170	tothrow
163	fall	81	to fall			40	to fall (drop)

	Semantically aligned list		Leipzig/Jakarta		Swadesh-100		Swadesh-200	
164	go	3	to go					
(119)	do, make	25	to do/make					
165	come	11	to come	66	come		23	to come
(64)	walk			65	walk		178	to walk
(147)	hunt						76	to hunt (game)
(150)	take	71	to take					
(119)	do, make	25	to do/make					
166	carry (bear)	70	to carry					
167	hold						74	hold (in hand)
168	road			85	path (road)		120	road
169	run	81	to run					
170	flow						52	toflow
(164)	go	3	to go					
(64)	walk			65	walk			
171	fly (v.)			64	fly		54	to fly
172	float						51	to float
173	swim			63	swim		159	to swim
174	sew						132	to sew
175	tie, bind	88	to tie				171	to tie
176	rope, cord	91	rope				122	rope
177	swell						158	to swell
178	spit						147	to spit
179	vomit						177	to vomit
180	turn over						175	to turn (veer)
181	thin (in dimension)						165	thin
182	narrow						101	narrow
183	small, little	91	small	15	small		140	small
(10)	little (quantity), few						46	few

Semantically aligned list		Leipzig/Jakarta		Swadesh-100		Swadesh-200	
184	short					134	short
(183)	small, little			15	small	140	small
(10)	little (quantity), few					46	few
(11)	some					146	some
184	near (adv.)					102	near
185	far (adv.)	23	far			41	far
186	long	78	long	14	long	92	long
187	large, big	32	big	13	big	11	big
188	much, many			10	many	95	many
189	wide, broad	96	wide			189	wide
190	thick (in dimension)					164	thick
191	heavy	71	heavy			71	heavy
192	new	53	new	96	new	104	new
193	sharp					133	sharp (knife)
194	blunt, dull					33	dull (knife)
195	dry			99	dry	32	dry (substance)
196	wet, damp					183	wet
197	rotten					123	rotten (log)
198	round			98	round		
199	full			95	full		
200	good	56	good	97	good	61	good
201	sweet	89	sweet				
202	smooth						
203	straight					155	straight
204	right, correct					117	right (correct)
205	right (side)					118	right (hand)
206	left (side)					87	left (hand)

	Semantically aligned list	Leipzig/Jakarta		Swadesh-100		Swadesh-200	
207	old	74	old			108	old
(187)	large, big	32	big	13	big	11	big
(197)	rotten					123	rotten (log)
(116)	smell (vb. trans)					141	tos mell (perceive odour)
208	bad					7	bad
209	dirty, soiled					29	dirty
210	black	42	black	91	black	14	black
211	bitter	28	bitter				
212	and					2	and
213	if					80	if
214	because					9	because
215	what?	50	what?	7	what?	184	what
216	when?					185	when
(213)	if					80	if
217	where?					186	where
218	how?					75	how
(215)	what?	50	what?	7	what?	184	what
(11)	some					146	some
219	who?	34	who?	6	who?	188	who
220	here					72	here
221	there					162	there
222	at					5	at
223	with					194	with (accompanying)
224	not	56	not	8	not	107	not
225	other					110	other

Note: the following connections indicated by CLICS have been ignored in elaborating the list: 'inside, in'–'if'; 'kill'–give'; 'year'–'leaf'; 'much, many'–'leaf'

[a] CLICS, the major source of information for organizing the list, makes a distinction between 'wood' and 'firewood', in this following its sources. The former has been chosen as it corresponds directly to the gloss in the Swadesh list. For 'firewood', a particularly close association to 'fire' is indicated, which should be checked if only a term for 'firewood' can be found for a given language one investigates.

[b] Here, CLICS distinguishes between 'head hair' and 'body hair', whereas both the Swadesh lists and the Leipzig/Jakarta list ask for 'hair' generally. Here, 'body hair' has been chosen for the list. If 'head hair' specifically is compared, connections may also be found with terms for 'head'.