Using Computational Criteria to Extract Large Swadesh Lists for Lexicostatistics

Johannes Dellert
Seminar für Sprachwissenschaft
Universität Tübingen
jdellert@sfs.uni-tuebingen.de

Armin Buch
Seminar für Sprachwissenschaft
Universität Tübingen
armin.buch@uni-tuebingen.de

***Abstract*—We propose a new method for empirically determining lists of basic concepts for the purpose of compiling extensive lexicostatistical databases. The idea is to approximate a notion of "swadeshness" formally and reproducibly without expert knowledge or bias, and being able to rank any number of concepts given enough data. Unlike previous approaches, our procedure indirectly measures both stability of concepts against lexical replacement, and their proneness to phenomena such as onomatopoesia and extensive borrowing. The method provides a fully automated way to generate customized Swadesh lists of any desired length, possibly adapted to a given geographical region. We apply the method to a large lexical database of Northern Eurasia, deriving a swadeshness ranking for more than 5,000 concepts expressed by German lemmas. We evaluate this ranking against existing shorter lists of basic concepts to validate the method, and give an English version of the 300 top concepts according to this ranking.**

I. Motivation

Lexical data is very valuable for studying the history and phylogeny of languages because it is comparatively easy to obtain, represent, and to compare across languages, but also unreliable since words can and often do cross language boundaries as loans. Swadesh lists (as originally devised in [1], [2], and extended or modified by many others[1]) are lists of stable concepts, i.e. concepts which are least likely to undergo lexical substitution or borrowing. A good Swadesh concept belongs to the basic vocabulary and is likely to be expressed by cognate words in phylogenetically related languages. An informal property summarizing how well a concept meets these criteria could be called its **swadeshness**.

The Concepticon [3] collects such lists of basic concepts. Typical Swadesh lists comprise around 200 items, and only a few specialized ones are longer. The longest list, the one employed in the World Loanword Database (WOLD) [4], with a core of 1460 and a total of 1814 concepts, is — as well as the one from the Intercontinental Dictionary Series (IDS) — an extended list of 'basic vocabulary' [5]. This covers just one of the two stated criteria for the inclusion of a concept. The trend in lexicostatistics has been to further reduce these lists, e.g. to concepts which most reliably represent language relationships (e.g. [6]). Current lexicostatistical databases with an emphasis on broad language coverage typically use short lists, e.g. a 40-concept list in ASJP [7]. The idea of using stochastic criteria on a very small number of highly stable concepts goes back at least to [8], who used a list of only 15 concepts to explore possible long-distance relationships in northern Eurasia.

However, in historical linguistics, such lists are considered insufficient to prove even established families, since 200 word pairs (not all of which are cognates) are not enough to detect regular sound correspondences.

To provide an empirical basis for collecting more extensive lexicostatistical databases, we derive a swadeshness ranking for a list of more than 5,000 concepts from weighted string distances on a very large cross-linguistic lexical database. As a proxy for swadeshness, we combine two relatively simple computational criteria which can be computed from any collection of bilingual wordlists with the same gloss language and a uniform phonetic transcription.

II. Data

Our work builds on a database which contains a total of 975,953 German translations for lemmas from 88 languages covering 20 different language families of Eurasia.

The version of the database we used contained more than 7,000 entries for at least one language from each of the following language families: Indo-European (English, German, Dutch, Swedish, French, Italian, Spanish, Portuguese, Russian, Polish, Latvian, Irish, Persian), Uralic (Finnish, Estonian, North Saami, Meadow Mari, Hungarian, Nenets, Nganasan), Turkic (Turkish), Mongolic (Mongolian), Sino-Tibetan (Mandarin Chinese), Austronesian (Indonesian), Semitic (Arabic), Kartvelian (Georgian), Japanese, and Basque.

To move beyond the size of existing concept lists, we first had to derive representations of a large number of candidate concepts. For this purpose, we inferred clusters of gloss lemmas from ratios of co-translations. This rough first version was then post-edited manually. The resulting sets of German lemmas serve as our representations of concepts. For example, the Swadesh concept EARTH/SOIL is represented by {*Boden, Erdboden, Erde, Grund, Land*}. Whenever the German lemmas are too polysemous or have homographs, our automated translation lookup procedure attempts to disambiguate by aggregating results across all the lemmas and annotations. The lookup of 5,088 concepts resulted in realizations across more than ten languages, which is the minimal number of languages we required to limit data sparseness issues.

[1] See http://concepticon.clld.org/contributions

Table I
INFORMATION CONTENT IN GERMAN *vergehen* "to pass" AND ARABIC *qatala* "to kill"

f	E	r	g	e	3	n
0.446	0.449	0.824	0.794	0.495	0.137	0.057

q	a	t	a	l	a
0.902	0.046	0.837	0.062	0.894	0.125

III. MEASURING INFORMATION CONTENT

For every language, either the orthography or additional pronunciation data contained in the database (e.g. for English, Danish, Japanese, Chinese, Persian) was converted into IPA and then into ASJP classes [9] using cascades of simple transducers, giving us a uniform phonetic representation for the entire database.

For each language and word class, n-gram models are used to quantify the information content of individual segments. Formally, if we write c_{abc}, c_{abX}, c_{Xbc}, c_{aXc} for the trigram and extended bigram counts, we define the **information content** of the segment c in its context $abcde$ as

$$I(abcde) := 1 - \max\left\{\frac{c_{abc}}{c_{abX}}, \frac{c_{bcd}}{c_{bXd}}, \frac{c_{cde}}{c_{Xde}}\right\}$$

In words, we use the minimum of the probabilities of not seeing c given the two segments before, the two segments after, and the immediate neighbors of c. This measure of information content serves a double purpose in our approach: It is used as a a generalized approach to stemming, and for normalizing word length in order to correct for effects of phoneme inventory size. Morphological material beyond the stem will have very low information content, because the inflection of citation forms (e.g. the 3rd person singular for verbs) will largely be predictable from the word class. Information content models inflectional and root-template morphology (as in Semitic languages) equally well (Table I).

We employ the information content in a modified string distance measure for Needleman-Wunsch style alignment of two strings a (length n) and b (length m):

$$M(i,j) := M(i-1,j-1) + w(a_i,b_j)\cdot s(a_i,b_j),$$

where the **combined information content** $s(a_i,b_j)$ is the quadratic mean of both surprisal scores:

$$s(a_i,b_j) := \sqrt{\frac{I(a_{i-2}\ldots a_{i+2})^2 + I(b_{j-2}\ldots b_{j+2})^2}{2}}$$

The quadratic mean is ideal because it encourages good alignment of segments with equally high information content, while not penalizing bad alignment of low-information segments too much, but strongly discouraging bad alignment of a high-information with a low-information segment. In the case of a gap, we define the combined surprisal score as the score of the non-gap segment, thereby discouraging the loss of important segments, and not penalizing the loss of less informative ones.

Table II
COMPARISON OF NORMALIZED WEIGHTED STRING DISTANCES

(English, German)	**Needleman-Wunsch**	**information-weighted**
(*measure*,*vermessen*)	0.258	0.175
(*reign*,*regieren*)	0.237	0.234
(*wash*, *waschen*)	0.317	0.290
(*pay*, *zahlen*)	0.478	0.600

The string distance measure is normalized through division by the sum of the combined surprisal scores in the best alignment of a and b:

$$d(a,b) := \frac{M(n,m)}{\sum_{i=1}^{align(a,b).length} s(align(a,b)[1][i], align(a,b)[2][i])}$$

Table II illustrates that cognate and non-cognate words between German an English are separated better by the information-weighted distance measure.

IV. MEASURING SWADESHNESS

A. Existing approaches

Already Swadesh's original paper [1, p. 457] suggests to use stability scores for creating longer Swadesh lists. Various stability measures have since been developed and applied to produce rankings of concepts.

Some measure stability directly as it is defined: as the extent to which a cognate class is preserved in a language family. This requires expert cognacy judgments [10]–[13]. In addition to cognacy, WOLD [14] quantifies and aggregates expert judgments on borrowing and historical data (earliest attestation of a word in a language). WOLD is unique in not only ranking the items of the Swadesh-100 or -200 list, but the entirety of its 1460 concepts. However, the method relies on expert data, and therefore does not easily scale up beyond WOLD's sample of 41 languages. Still, the well-distributed language sample means the ranking would not change too much with the addition of more data.

For larger and more recent data sets, expert judgments are not readily available. In one line of research [6], [9], [15], discrete cognacy judgments are replaced by the average string similarity between all realizations of a concept in a language family. State of the art is LDND [16], i.e. Levenshtein distance normalized for length and divided by the expected chance similarity between words of the two languages in question. The latter is calculated as the average similarity of words with different meanings. A recent suggestion [17] further abstracts away from cognacy by measuring the diversity of realizations of a concept (within each language family) as the self-entropy of these realizations seen as a bag of n-grams: A short-tailed distribution of n-grams indicates fewer lexical substitutions (and less phonemic diversity), whereas a heavy-tailed distribution reflects the existence of many cognate classes for this concept.

Automated approaches can easily be applied to thousands of concepts if dictionary data is available. Such measures of swadeshness exclusively rely on massively cross-linguistic

data, avoiding any expert bias e.g. in favor of a particular language family. At the same time, if such measures are clearly defined and reproducible, they can be employed e.g. to determine good basic concept lists for new geographical areas.

All methods referenced here are scalable a priori. However, with the exception of WOLD, they have so far merely been applied to rank and subsequently narrow down lists which were short to start with (100 items of ASJP or Swadesh's 100). As our method is based on semi-automatic inference of concepts from raw data, we are the first to also consider concepts that were not chosen by experts.

Also, all existing methods require language families to be defined, such that a match (non-match) within a family is evidence of stability (replacement). The reverse argument, that a match across families is evidence of borrowing, is mostly neglected, since cross-family borrowings are both rare in the Swadesh-100 core vocabulary and difficult to detect. When ranking more concepts, neglecting this negative evidence will result in high scores for onomatopoeia (CUCKOO) and *Wanderwörter* (COFFEE, GAS). Our measure combines both types of evidence: In addition to abstracting over discrete cognacy judgments, we will abstract over discrete families, by taking very close languages stronger into account as evidence of the first type, and very distant languages for evidence of the second type.

B. Our measure

Since swadeshness is an inherent property of a concept which we cannot observe directly, our approach attempts to determine swadeshness more indirectly by a combination of two measurable indicators. The first measure implements a bias in favor of basic vocabulary, building on the assumption that more basic concepts will tend to have shorter realizations. To compare word lengths cross-linguistically, we simply add up segment-wise information content over the ASJP strings. Averaging over all languages in which realizations for a concept c are available, we derive the average information content $inf(c)$.

The second measure quantifies stability over time without building on expert cognacy judgments. Instead of applying automated cognate detection, our approach is to measure how well the distances between realizations of each concept represent overall language distance. We compute concept realization distances for every concept and language pair based on the Needleman-Wunsch algorithm, with segment distances trained separately for each language pair, just as implemented in LexStat [18]. Furthermore, we derive global language distances using the dER measure [19], which controls for the influence of random similarity by computing similarity values between all pairs of realizations from the two languages across all concepts, and then deriving an aggregated similarity measure from the ranking of pairs of the same meaning in the ranking of similarity values for different meanings. For the global distances, we restricted ourselves to the realizations of the top-50 concepts from the Holman ranking [6]. 1,250 pairs are enough for a good estimate, while avoiding problems with massive loans in less basic vocabulary.

Our local-global distance correlation $lgc(c)$ is the average Pearson correlation between concept-specific distances and global distances over 10 balanced samples of language pairs. Balanced sampling is necessary because the majority of language pairs is unrelated, leading to a high density of points in the high global score range. The balanced samples are drawn by uniform sampling in the global score dimension, and selecting the nearest point for each sampled global score value.

By using the correlation, we penalize similar realizations in unrelated languages (e.g. English *door* and Japanese *doa*, a borrowing) as well as dissimilar realizations in closely related languages (e.g. German *Körper* and Dutch *lichaam* "body", where the former is a Latin loan).

Our ranking is then based on the **swadeshness score** $sc(c) := inf(c) - 3 \cdot lgc(c)$, a simple linear combination with a single empirically determined parameter, which we optimized for high coverage of the Swadesh-207 list (the union of the 100- and 200-item lists). Such a simple model for combining the two measures minimizes the risk of overfitting.

V. RESULTS AND DISCUSSION

English translations of the top 300 concepts in our ranking are given as an appendix. A comparison to non-ranked lists is possible in terms of coverage: Table III shows how much of Swadesh's 207-item list, the ASJP list, the Leipzig-Jakarta list [20] and WOLD's concept list is covered by the top n concepts of our ranking. Some concepts are not covered at all by the ranking procedure, because lookup in the dictionary database is unsuccessful. This is the case when either the German lemma is very ambiguous, making the concept indistinguishable from another concept; or the lemma is unambiguous, but too specialized to occur in the database. If these lexical lookup difficulties can be overcome (e.g. by more extensive manual annotation), the missing concepts fare quite favorably:

- Translations of the concept IT are mapped to German *es* in the automatically inferred set of lemmas *{es,ihm,ihn,ihr}* because of the case ambiguity of *es*. We manually mapped this concept to the WOLD concept HIM OR HER (rank 13), causing this apparent gap in the coverage of Swadesh-207 and Leipzig-Jakarta to vanish.
- We furthermore miss the Leipzig-Jakarta concept TO DO, which German does not systematically distinguish from TO MAKE (rank 141 for *machen*).

Comparing our ranking to other published rankings on the ASJP-100 list, we get Spearman rank correlation values of 0.276 [6] and 0.468 [17], which indicates we are catching a similar signal even among the most stable concepts, even though this was not our main goal.

More importantly, 95 of 207 Swadesh concepts are also in our top-207 list, which we consider a rather promising result given that they have been sifted out from over 5,000 candidates. This makes our ranking a valuable empirical source of high-swadeshness concepts.

Table III
COVERAGE

list	covered in first...						total
	100	207	500	1000	2000	5088	
Dolgopolsky	11	13	14	14	15	15	15
Swadesh-207	57	95	150	184	204	206	207
ASJP	41	60	83	98	99	100	100
Leipzig-Jakarta	38	57	77	93	97	98	100
WOLD	93	183	391	665	1023	1385	1814

ACKNOWLEDGMENTS

This work was supported by the ERC Advanced Grant 324246 "EVOLAEMP", which is gratefully acknowledged. We would like to thank Søren Wichmann for essential feedback, and the many members of the EVOLAEMP team who contributed some additional data to the database.

REFERENCES

[1] M. Swadesh, "Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos," *Proceedings of the American Philosophical Society*, vol. 96, no. 4, pp. pp. 452–463, 1952.

[2] ——, "Towards Greater Accuracy in Lexicostatistic Dating," *International Journal of American Linguistics*, no. 2, pp. 121–137.

[3] J.-M. List, M. Cysouw, and R. Forkel, Eds., *Concepticon*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2015. [Online]. Available: http://concepticon.clld.org/

[4] M. Haspelmath and U. Tadmor, Eds., *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2009. [Online]. Available: http://wold.clld.org/

[5] C. D. Buck, *A dictionary of selected synonyms in the principal Indo-European languages: A contribution to the history of ideas*. Chicago: University of Chicago Press, 1949.

[6] E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, and D. Bakker, "Explorations in automated language classification," *Folia Linguistica*, vol. 42, no. 3-4, pp. 331–354, 2008.

[7] S. Wichmann, A. Müller, A. Wett, V. Velupillai, J. Bischoffberger, C. H. Brown, E. W. Holman, S. Sauppe, Z. Molochieva, P. Brown, H. Hammarström, O. Belyaev, J.-M. List, D. Bakker, D. Egorov, M. Urban, R. Mailhammer, A. Carrizo, M. S. Dryer, E. Korovina, D. Beck, H. Geyer, P. Epps, A. Grant, and P. Valenzuela, "The ASJP Database (version 16)," 2013.

[8] A. B. Dolgopolsky, "A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia," *Typology, Relationship, and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists, eds. VV Shevoroshkin and TL Markey. Karoma, Ann Arbor, MI*, pp. 27–50, 1986.

[9] C. H. Brown, E. W. Holman, S. Wichmann, and V. Velupillai, "Automated classification of the world's languages: A description of the method and preliminary results," *Language Typology and Universals*, vol. 61, no. 4, pp. 285–308, 2008.

[10] I. Dyen, A. James, and J. Cole, "Language divergence and estimated word retention rate," *Language*, pp. 150–171, 1967.

[11] R. L. Oswalt, "Towards the construction of a standard lexicostatistic list," *Anthropological Linguistics*, pp. 421–434, 1971.

[12] J. B. Kruskal, I. Dyen, and P. Black, "Some results from the vocabulary method of reconstructing language trees," *Lexicostatistics in genetic linguistics*, pp. 30–55, 1973.

[13] C. Peust, "Towards establishing a new basic vocabulary list (Swadesh list)," 2013. [Online]. Available: http://www.peust.de/peustBasicVocabularyList.pdf

[14] U. Tadmor, "Loanwords in the world's languages: Findings and results," in *Loanwords in the world's languages. A comparative handbook*, M. Haspelmath and U. Tadmor, Eds., 2009, pp. 55–75.

[15] F. Petroni and M. Serva, "Automated Word Stability and Language Phylogeny," *Journal of Quantitative Linguistics*, vol. 18, no. 1, pp. 53–62, 2011.

[16] E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, D. Bakker *et al.*, "Advances in automated language classification," 2008.

to be, what, to give, I, one, thou, water, to arrive, to put, to drink, to go, to come, him/her, they, she, to take, to walk, he, hand, you, this, in, two, and, day, if, blood, at, mouth, to carry, month, river, to say, with, side, we, ten, to burn (intransitive), to live, arm, tooth, soil, louse, who, person, one (pronoun), to melt, skin, ice, to go away, to see, by means of, to float, grease/fat, is, foot, six, that, to sleep, hair, to listen, to become, to travel, to fetch, there, to buy, to tie, to hear, name, head, here, man, son, fingernail, to bring, house, to water, to know, since, night, where, end (temporal), good, road, ashes, us, branch, dog, not, path, to wear, to cry, stone, to stand up, like that, to sew, to fall, thee, direction, them, five, fish, like this, to cover, to divide, under, to get, his/her, fire, to pound, a, thread, bone, to hold, wood, to remain, throat, to blow, snow, thy, place, eye, to smelt, to call (give a name), moon, to step, to fly, new, land, to read, in order to, big, to cease, salt, to boil, via (direction), woman, to make, to open, to you (pl.), whom, to plait, nest, breast, strength, but, to mow, to cook, to flow, too, to pursue (a business), whether, neck, leg, mother, old age, to seem, beech, can, thither, word, to wash, nose, cold, towards, work, ear, old man, three, to pour, a hundred, to cut, horn, to throw, to share, to weave, father, wool, to stand, or, to be allowed, grass, door, to milk, to dig, bile, my, peel/husk, god, mind, to have, to wipe, warm, while, language, to smell (intransitive), long, to bite, to burn (transitive), me, to suck, to find, chest, tongue, other, to sink, cow, meadow, to grow, to roast, edge, sea, country, to kick, to finish, shaft, half, than (comparison), seat, to tread, lake, horse, wet, to die, tree, to lick, arrow, mountain, to look, in (time span), four, clay, under (direction), bad, to dry (intransitive), to hang up, to swim, to elevate, beeswax, town square, to thee, to abandon, to eat, leaf, to incline, smoke, to recognize, hither, topic, pain, to plough, stick, through, to choose, bark, hedge, time, bar, to fill, page, above, let, time (occasion), top, as (comparison), handle, green (n.), to row, price, meat, law (subject), out (direction), to feel, more, moss, to hit, to begin, wide, deep, to want, steady, winter, to rinse, amount, to move, self, wind, to pull, for, sail, belt, hole, to suit, to tear, to sell, to preserve, age, saddle, ago, to sting.

Figure 1. The first 300 concepts in our ranking. Swadesh concepts (117/207) in blue and violet, ASJP concepts (72/100) in violet and red (ASJP only differs in having *path* instead of *road*). Sparse data (< 20 languages) in gray.

[17] T. Rama and L. Borin, "N-Gram Approaches to the Historical Dynamics of Basic Vocabulary," *Journal of Quantitative Linguistics*, vol. 21, no. 1, pp. 50–64, 2014.

[18] J.-M. List, "LexStat. Automatic detection of cognates in multilingual wordlists," in *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, Stroudsburg, 2012, pp. 117–125.

[19] G. Jäger, "Phylogenetic Inference from Word Lists Using Weighted Alignment with Empirically Determined Weights," *Language Dynamics and Change*, vol. 3, no. 2, pp. 245–291, 2013.

[20] M. Haspelmath and U. Tadmor, *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter, 2009.