A new approach to concept basicness and stability as a window to the robustness of concept list rankings

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

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

Abstract

Based on a recently published large-scale lexicostatistical database, we rank 1,016 concepts by their suitability for inclusion in Swadesh-style lists of basic stable concepts. For this, we define separate measures of basicness and stability. Basicness in the sense of morphological simplicity is measured based on information content, a generalization of word length which corrects for distorting effects of phoneme inventory sizes, phonotactics and non-stem morphemes in dictionary forms. Stability against replacement by semantic shift or borrowing is measured by sampling independent language pairs, and correlating the distances between the forms for the concept with the overall language distances. In order to determine the relative importance of basicness and stability, we optimize our combination of the two partial measures towards similarity with existing lists. A comparison with and among existing rankings suggests that concept rankings are highly data-dependent and therefore less well-grounded than previously assumed. To explore this issue, we evaluate the robustness of our ranking against language pair resampling, allowing us to assess how much volatility can be expected, and showing that only about half of the concepts on a list based on our ranking can safely be assumed to belong on the list independently of the data.
Keywords

basic vocabulary – Swadesh list – stable concepts – information-weighted sequence alignment – concept stability – computational historical linguistics

1 Introduction

Lists of basic vocabulary have always been a cornerstone of linguistic fieldwork, as they provide the raw data for many research questions in historical linguistics. Field linguists prefer basic concepts that are realized by a single morpheme in many languages because they tend to be easier to elicit, and from the lexicostatistical point of view they are preferable because single-morpheme words are more likely to provide independent data points.

The second factor increasing the value of a concept to a historical linguist is stability against lexical replacement, because the word forms for stable concepts carry a deep historical signal providing hints about the language’s ancestry. Swadesh (1952) pioneered the idea of defining a short list containing only the most stable basic concepts, and using them to date the latest common ancestor of a language pair based on an assumed fixed replacement rate.

While this assumption has long been found to be incorrect (e.g. Bergsland and Vogt, 1962), computational approaches to historical linguistics still rely on short concept lists that are very similar to Swadesh’s lists, although they have since undergone a number of revisions. The preliminary nature of every ‘Swadesh list’ is evident in Swadesh’s own publications (Swadesh, 1952, 1955, etc.). Once published, the lists tend to be taken as securely established and reused across lexical data collection projects. However, the reproducibility and robustness of these concept lists is seldom addressed. How certain are we that a concept included in a list of 100 basic stable concepts is actually among the 100 most valuable concepts for doing historical linguistics in a given region?

Reproducing the process by which other authors arrived at their concept lists (even when using the same data) is obviously difficult when their method is informal or undocumented, but even formal, quantitative criteria cannot guarantee reproducibility when relying on expert knowledge, because experts’ assessments will differ (e.g. on whether a word is a loanword or not). The obvious and commonly used remedy to this issue is to use fully automated data-driven methods, which guarantee to always produce the same result when run multiple times on the same dataset.
A lack of robustness, i.e., a high volatility of the outcome under changes in the input data, can be an issue even for fully specified and perfectly reproducible data-driven methods. To get a first impression of the scale of the problem, we take a closer look at the very popular Leipzig-Jakarta list\(^1\) (Tadmor, 2009), derived from the World Loanword Database (\textsc{WOLD})\(^2\) (Hapnelmath and Tadmor, 2009). Its scope is the Loanword Typology meaning list with 1460 concepts, revised from the Intercontinental Dictionary Series\(^3\) (Key and Comrie, 2015), which in turn is a revision of Buck (1949).\(^4\) This list has thus come a long way from its initial focus on Indo-European languages, but it is unclear how much of that bias might remain. The possibility of selection biases induced by imbalances in the language sample the lists are induced from is an issue for any concept list. In general, a concept list for historical linguistics inferred predominantly from languages of a given region might not be an optimal list for another region.

\textsc{WOLD} contains quantified expert judgments on borrowing, age and morphological complexity, and reports per-concept averages over its sample of 41 languages. Averaging of values is intended to counter any bias the experts might have had. Tadmor (2009) scores each concept by the product of the three measures and a concept’s representation ratio in \textsc{WOLD}’s language sample.

A simple way to test the reproducibility and robustness of the resulting Leipzig-Jakarta (LJ) list is to attempt reproducing it from a more recent (2016/09/29) snapshot of \textsc{WOLD}. We recalculated Tadmor’s ranking for the \textsc{WOLD} concepts based on the scores now reported. Perhaps surprisingly, the list inferred from the newer data differs from the published list by 16 out of 100 items (Table 1). Also, the rankings of the 84 overlap items differ widely: The Spearman rank correlation between the values reported in Tadmor (2009) and our reproduction is only 0.68 (95% CI: [0.56, 0.78]). The details of this replica experiment are part of the online supplementary materials for this article.\(^5\)

The unreliability of the Leipzig-Jakarta list is an issue: The IELex database (Bouckaert et al., 2012),\(^6\) which previously used the joint Swadesh-100 (Swadesh, 1955) and Swadesh-200 (1952) list,\(^7\) has been extended to include the 18

---

\(^1\) Entry in Concepticon (List et al., 2015), an online catalogue of concept lists: http://concepticon.clld.org/contributions/Tadmor-2009-100.
\(^2\) http://concepticon.clld.org/contributions/Haspelmath-2009-1460.
\(^3\) http://concepticon.clld.org/contributions/Key-2016-1310.
\(^4\) http://concepticon.clld.org/contributions/Buck-1949-1108.
\(^5\) Available at https://figshare.com/s/c9cf935761492899d86.
\(^6\) Available at http://ielex.mpi.nl/.
\(^7\) Concepticon lists five attempts to merge the lists, all yielding 207 items, but with slightly differing semantics.
| Additional items on new list | Missing items of old list |
|-----------------------------|--------------------------|
| AT we CHILD-IN-LAW knee dog blow (of wind) | KNEE DOG BLOW (OF WIND) |
| UDDER weave FIREWOOD | RED WIDE ASH |
| BE pus SHOULDER | THIGH GRIND OLD |
| GRASP EDGE CHEEK | SUCK SEE SWEET |
| THROW BELLY MOUNTAIN | TAKE CRY SALT |
| HAVE | HEAR |

items missing from the Leipzig-Jakarta list. Five of these would not be on the list under our recompilation of the Leipzig-Jakarta list, and five others would be added instead. That is to say, the data-driven and thus seemingly objective choice of which concepts to include is not justified any longer with a newer version of the data, which demonstrates that there is a robustness issue.

The lesson is that even lists which were deterministically derived from the data (such as Leipzig-Jakarta from WOLD’s scores) can vary a lot with the data they are inferred from. On some level, the inclusion or exclusion of a concept will always be subject to chance, and both creators and users of concept rankings should be aware of their volatile character.

In this article, we present two new separate measures of basicness and stability, which we combine into a new method for ranking concepts. Reproducibility is ensured by fully automatizing the process from the level of word forms to the final combined score. We test our method on the NorthEuraLex database9 (Dellert et al., 2017) which consists of IPA representations of the words for 1016 concepts in 107 languages covering 21 language families. The IPA representations are derived in a semi-automated fashion from standard orthographies, and are therefore neither fully phonemic nor fully phonetic representations with some inaccuracies.

The last section of this article is devoted entirely to evaluating the volatility of our measure against changes to the dataset. Using resampling to derive a large number of rankings from the same dataset, we show that our measure

---

8 Also, IELex retains the woods = forest and child (young human), although Tadmor (2009) clearly refers to wood and child (descendant).

9 Version 0.9, available at www.northeuralex.org.

10 With a few gaps; at least one form is given for 97.9 % of all concept-language pairs.
can provide an approximation to good concept lists of any length, but is not fine-grained enough to make binary inclusion decisions between concepts of medium basicness.

2 Combining basicness and stability

It would be difficult to define a single criterion measuring both the basicness and the stability of a concept, as can be demonstrated using examples from the NorthEuraLex database. A concept like TEA is basic and expressed by a very short word form in almost every language, but this word is a wanderwort and therefore a borrowing in almost any language. In contrast, words for SIXTY typically contain two inherited morphemes, both of which are quite stable against borrowing and therefore often contain more historical signal than the forms for SIX alone, but SIX is still preferred in concept lists because it is more basic. We therefore attempt to capture basicness and stability by separate measures, which we then combine to model the criteria which implicitly underlie existing lists.

The first measure models basicness based on the assumption that more basic concepts will tend to have shorter realizations because they will be expressed more frequently by single morphemes. The second measure quantifies stability over time without building on expert cognacy judgments. Instead of attempting to apply automated cognate detection, our approach is to measure how well the distances between forms for each concept reflect the overall language distances.

2.1 Basicness measure

To compare lengths of word forms in a cross-linguistically valid way, it will not suffice to simply compare the number of segments in the IPA representations. Some languages have richer phoneme inventories than others. Some languages have short stems that are either monosyllabic (Hungarian, Thai) or consist of a single consonant cluster (Georgian), whereas in other languages (Spanish, Finnish) almost every stem is at least disyllabic. Also, the dictionary forms included in lexical databases often feature infinitive endings or other morphological material that only includes information about the category, but in no way informs us about the complexity of the stem.

Another issue for a good definition of word length based on phonetic descriptions is the presence of allophones, which often produce segments that are determined entirely by their contexts. For instance, in the NorthEuraLex IPA encoding, where length is expressed by copying segments, the presence of
the phoneme [u] in a German word form predicts that either the preceding or the following segment will also be [u], since tense vowels in German are always long.

To correct for the distorting effect of these phenomena, we employ a segment-wise information content model, and add up the information content across segments to compute an adjusted word length. For each segment, the information content measures the predictability of this segment based on a five-segment context. This predictability is modeled by comparing counts of trigrams (sequences of three consecutive segments, e.g. [seg] or [gma] in the pronunciation of segments) with the counts of gappy trigrams or extended bigrams (where one element can be an arbitrary symbol X, e.g. [sXg] and [gmX] for our trigram examples). If almost every time we encounter [gmX], the full trigram is actually [gma], this means that [a] is predictable from the preceding segments, and should therefore have low information content. To have a full five-segment context even for peripheral segments of the string, we add the word boundary symbol # twice before the beginning and after the end of each string, both for deriving the trigram and extended bigram counts and for defining the positions $a_{−1}$, $a_0$, $a_{k+1}$, and $a_{k+2}$ in a string $a$ of length $k$.

Formally, if we write $c_{abc}$, $c_{abX}$, $c_{Xbc}$, $c_{axc}$ for the trigram and extended bigram counts extracted from all word forms of a language $L$, we can define the information content of a segment $c$ in its five-segment context $abcde$ (inside a word of any length) as

$$I_L(c, [ab_de]) := -\log \left( \frac{c_{abc} + c_{bcd} + c_{cde}}{c_{abX} + c_{bXd} + c_{Xde}} \right)$$

We thus add up the counts of the trigrams consisting of $c$ and the two segments before, the two segments after, and the immediate neighbors of $c$, and divide this count by the sum of the corresponding extended bigram counts. Since each extended bigram count is just the sum of the counts for all the trigrams it represents, the fractions for all possible values of $c$ in a fixed context $[ab_de]$ will sum up to 1, which means that they can be taken to represent a probability distribution. Among the options to define such a distribution for each context, summing up the observations of all three trigrams in the context has the advantage of ensuring that more frequent contexts count more, while still letting all three contexts influence the probability, which somewhat alleviates problems otherwise caused by small bigram counts for one of the contexts.

However, if we compute the information content based on actually observed trigram counts, this can lead to much too strong inference from sparse data. An extreme case of this is when for some context all three extended bigrams
only occur once. This would cause the quotient in the definition to be identical to 1, i.e., we would infer complete predictability based on a single instance. As in most models based on empirical observation counts, some smoothing is therefore necessary. Smoothing can be defined as re-assigning probability to unseen events in order to limit the impact of single observations. An easy way to smooth our distributions is to uniformly add a small number of pseudo-observations to all (including unseen) trigrams. To decide how many observations to add, it is necessary to find a balance between limiting the influence of small counts and weakening the differences too much. We heuristically determined that this balance was achieved well enough by uniformly distributing an additional 20% of the total number of observations over all trigrams. For instance, in the model without smoothing, the trigram [dɹæ] in *drag* is the only instance of the extended bigram [dXæ] in the English part of NorthEuraLex. This would cause the model to assume that [ɹ] is completely determined by its context [d_æ], based on the assumption that the database completely represents the phonotactic possibilities of English. While there are many words justifying the assumption of a high probability (*drab*, *dragon*, *drastic*), the presence of a word like *dwang* shows that we need to model different possibilities as well. Smoothing does just that by assigning some probability to unseen events. After smoothing, the observation count for [dXæ] is 1.305, and [dɹæ] is assumed to only represent 76.6% of all instances of [dXæ].

While smoothing is necessary to handle the cases in which observation counts are low, the usefulness of the information content measure shows much more clearly when considering a case where we do have sufficient information. As an illustrating example, we take the information content $I_{deu}(n,[eə_##])$ of the segment [n] in the context [eə_##] in German words, e.g. in the pronunciation of the verb *vergehen* ‘to pass.’ The relevant smoothed counts are $c_{con} = 17.001$, $c_{eəX} = 24.124$, $c_{oəX} = 604.924$, $c_{n##} = 426.001$, and $c_{X##} = 1219.324$. The high numbers of observations for the right context (i.e. the word endings [ən] and [n]) mean that the information content will depend mainly on the probability for a German word to end in [ən]. The fraction evaluates to $\frac{811.003}{1848.372} = 0.439$, which we interpret as our inferred probability for [n] to occur in the context [eə_##], i.e. at the end of a word after [eə].

Coming to the final step in our definition of information content, the negative logarithm is a standard technique from information theory for converting the probability of an event into its self-information measured in bits or nats (depending on the logarithm used). For our example, we get $I_{deu}(n,[eə_##]) = 0.824$ nats of information, which is equivalent to 1.189 bits. 0.918 bits or 77% of this information can be interpreted to encode whether a German word is a verb (33.5% of the German forms in the database) or not, and only very little
additional information is needed because virtually all German infinitives, and few other words in the database, end in -en.

Morphological material beyond the stem will typically be low in information content, because the inflection of citation forms (e.g. the infinitive endings of verbs) will occur very frequently. For instance, expressing the relative information content using shades from white to black, the information content of the segments in vergehen is [fɛŋɡen], i.e., the infinitive ending disappears almost completely, and the verbal prefix ver- only contains about 75% of the information of the root geh-. This root is correctly interpreted to contain only two high-information segments, as the first [e] can be inferred from the fact that the sound [e] only occurs as a long vowel [ee] in German (see Table 2 for exact numbers).

Interestingly, this definition of information content covers affixal and root-template morphology equally well. For Arabic haraba ‘to flee,’ we get the desired result [haraba] where the consonants contain 77% of the information (numbers in Table 2). Our simple information model is thus powerful enough even to detect the three-consonant root structure of Semitic languages.

Based on our information content model for word forms, we can define the average information content \( \text{inf}(C) \) of a concept \( C \) as the average of the language-specific information contents of all available forms for \( C \). The average information content will serve as our measure of (non-)basicness.

| TABLE 2 | Information content (in nats) in German vergehen ‘to pass’ and Arabic haraba ‘to flee’ |
|----------|---------------------------------|
| [f]      | [ɛ]    | [v] | [g] | [e] | [e] | [ə] | [n] |
| 2.213    | 0.963  | 0.558 | 2.279 | 0.372 | 2.369 | 0.487 | 0.824 |
| [h]      | [a]    | [r] | [a] | [b] | [a] |
| 4.416    | 0.943  | 2.291 | 0.988 | 2.741 | 0.932 |

2.2 Information-Weighted Sequence Alignment

For our stability measure, we first need to define a way to compute distances between phonetic forms in the database. State-of-the-art methods for this task in the context of cognate detection are based on weighted sequence alignment. As was the case for word length, we use the language-specific information models to generalize an existing method to be more language-independent and to produce scores that are less sensitive to the distorting effects of repeating cognate or non-cognate morphological material.

For weighted sequence alignment, we build on a global phoneme similarity matrix derived from the database. As in the estimation procedure presented
by List (2012), our algorithm iteratively refines a candidate set of cognacy pairs. In the first iteration, these candidates are filtered out based on edit distance (Levenshtein, 1965) normalized by division through the length of the longer string, and on the Needleman-Wunsch scores (Needleman and Wunsch, 1970) of the previous run during subsequent iterations. The re-estimated scores are pointwise mutual information scores based on comparing the correspondences extracted from optimal alignments of the current cognate candidate pairs with correspondences occurring in alignments of a large set of random word pairs. Appendix A1 provides a more formal description of the estimation process.

Segment distances and information content are then combined into a modified string distance measure for Needleman-Wunsch-style alignment of two strings $u$ and $v$, where the segment similarity scores $w(a, b)$ are weighted by the combined information content $I_{L_a, L_b}^2(a, b)$, defined as the quadratic mean of both information contents $I_{L_a}(a, x_a)$ and $I_{L_b}(b, x_b)$ of the segments $a$ and $b$ in their respective languages $(L_a, L_b)$ and five-segment contexts $(x_a, x_b)$. Any generalized mean could serve to encourage good alignment of informative segments while not penalizing bad alignment of low-information segments. Among standard choices of generalized means, the quadratic mean was heuristically determined to be preferable because it discourages bad alignment of a high-information segment with a low-information one, a desired effect to prevent endings from being matched with stems. In the case of insertions and deletions (a gap in the alignment), we define the weight as the information content of the non-gap segment (i.e. the quadratic mean of the information with itself), thereby discouraging the loss of important segments and not penalizing the loss of less informative ones.

As a name for the new alignment method, we propose the name **Information-Weighted Sequence Alignment (IWSA)**. The method is illustrated by two examples in Figure 1, and the complete algorithm for computing the form similarity scores $sc(u, v)$ for word forms $u$ and $v$ is given in Appendix A2. As in List (2012), $sc(u, v)$ is normalized through division by the average self-similarity of both word forms according to the respective information model for each language, leading to the following definition of **Information-Weighted Distance (IWD)**:

$$d(u, v) := 1 - \frac{2 \cdot sc(u, v)}{sc(u, u) + sc(v, v)}$$

The same normalization applied to the Needleman-Wunsch score without information weighting (i.e. $I_{L_a, L_b}^2(a, b) = 1$ for all segment and language pairs)
will be called **Needleman-Wunsch Distance (NWD)**. The \(d(u, v)\) computed in this way are not necessarily in the range \([0, 1]\), but can be slightly below 0 and above 1 in extreme cases, e.g., if most of the scores \(w(a, b)\) are negative and both forms have high information density. This will not be an issue for our application, however.

To evaluate whether information weighting helps to better separate cognate from non-cognate word pairs, we applied both NWD and IWD to the IELex-based cognate pair testset provided in the supplementary materials of Dellert (2017). On this testset of 100,156 word pairs from Indo-European languages included in NorthEuraLex, a binary cognacy classification makes it possible to derive precision and recall numbers for any choice of a threshold value to separate cognate from non-cognate form pairs. To compare the methods in a threshold-independent fashion, we use the average precision, which is defined as the area under the precision-recall curve generated by considering all possible threshold values. For List’s LexStat distance as implemented in the LingPy system, we receive an average precision of 0.728 (95% CI: [0.723, 0.732], determined from 1,000 bootstrap samples). Our NWD implementation turned out to perform very similarly at 0.731 (95% CI: [0.726, 0.735]), although, unlike LexStat, the variant we present here does not infer separate phoneme similarity matrices for each language pair. Our implementation of NWD thus shows performance in the range of the current state of the art. For IWD, average precision rises to 0.747 (95% CI: [0.742, 0.752]), showing that information weighting does significantly improve the quality of phonetic sequence alignment.

For word pairs from the same genus, the already very good performance of LexStat at an average precision of 0.972 (95% CI: [0.969, 0.974]) does not leave any room for improvement for IWD at 0.974 (95% CI: [0.971, 0.976]). In contrast, the difference for word pairs from different genera is quite marked, with IWD at an average precision of 0.567 (95% CI: [0.559, 0.575]) against LexStat at 0.514 (95% CI: [0.506, 0.521]). The measurable advantage of IWD is therefore mostly due to improvements in cognacy decisions among distant relatives, where all methods still perform rather poorly. But at the same time, IWD is also more reliable in separating cognate and non-cognate forms in some pairs of closely related languages. Table 3 illustrates this with a few English and German examples taken from the testset.
2.3 Stability measure

Based on a model that provides high-quality form distances, we can now define our stability measure. The idea is to determine how well the information-weighted distances between the words for a concept mirror the overall language distances. Such a measure will not capture every relevant aspect of stability. For instance, distances between forms for concepts that are unstable against internal replacement will still tend to correspond rather well with the overall language distances, especially for more closely related languages.

However, a single measure that quantifies how closely form distances are aligned with language distances will allow us to summarize effects generated by several independent phenomena which factor into the notion of stability. For instance, distances between forms for concepts that are unstable against internal replacement will still tend to correspond rather well with the overall language distances, especially for more closely related languages.

If overall language distances are defined on the same scale as the form distances, such a measure can be defined in a very straightforward way by the correlation between form distances and language distances for a sample of language pairs. In this way, we do not only capture whether the words in related languages are more similar, while they should not be similar at all (except by chance) for language pairs from different families. Within a family, similarity of inherited forms should decrease with the age of the latest common ancestor of the language pair. On the other hand, if a concept is borrowed between two distantly related or unrelated languages, this will lead to a low form-pair distance for a high-distance pair of languages.

If overall language distances are defined on the same scale as the form distances, such a measure can be defined in a very straightforward way by the correlation between form distances and language distances for a sample of language pairs. In this way, we do not only capture whether the words in related languages are more similar, but also 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).

To determine the language distance for each language pair, we simply average the information-weighted form distances for all concepts where forms
in both languages are available, causing form and language distances to be on similar scales.

Directly computing the correlation between all available form distance and language distance pairs is problematic because the language pairs will not constitute independent samples. The samples can be made much less dependent if we only use each language twice in a language pair sample, and in such a way that a language pair \((L_1, L_3)\) is never drawn if \((L_1, L_2)\) and \((L_2, L_3)\) are already in the sample for some language \(L_2\). More generally, we want to avoid cycles of any size in the graph over languages defined by the sampled language pairs. This is equivalent to sampling an arbitrary tree-shaped graph spanning the sampled languages.

A small subset of these tree-shaped sample graphs can be generated by randomly permuting the list of languages, and interpreting the permutation as describing a chain of language pairs. The maximum sample size is then one less than the number of languages in which data for a given concept is present, giving us more than a hundred samples for all but a handful of concepts. This small number of samples will suffice for our purposes, but depending on the structure of the language family described by the data, more sophisticated sampling techniques might be needed to ensure that a number of closely related language pairs will occur in each language pair sample.

Our stability measure for a concept, which we call \textbf{local-global distance correlation (lgc)}, is defined as the Pearson correlation of form and language distances averaged across 100 of these language chain samples, making sure that all our scores are near the middle of the range that would be generated by different chain samples. In Section 4, we analyze the volatility of the rankings computed from single chain samples to assess how much the ranking depends on the data.

3 Optimization and quantitative evaluation

3.1 \textit{Defining a mixed measure to model existing lists}

The goal of our measure is to capture the signal behind the inclusion decisions for existing concept lists, which require the concepts to be both basic and stable against replacement. After developing approximate measures of basicness and stability in Section 2, in this section we try to model both criteria at the same time by combining both measures. We decide how to combine them by maximizing the number of concepts on existing shorter lists that end up in top positions in the ranking by our combined measure. For this purpose, we use four widespread lists that arose from a preselection of concepts given
A new approach to concept basicness and stability

169

Language Dynamics and Change 8 (2018) 157–181

Table 4

Optimizing $\lambda$ by minimizing $\mu(\lambda, K)$. $\lambda_{opt}$ is the optimal convex coefficient for each list, $\lambda = 0.82$ the global optimum. All numbers involving $lgc$ with 95% confidence intervals derived from 600 recomputations of $lgc$.

| List  | $\lambda_{opt}$ for list | $m_{\lambda_{opt}, K}$ | $m_{0.82, K}$ | $m_{1, K} (lgc)$ | $inf$ |
|-------|---------------------------|------------------------|---------------|-----------------|------|
| ABVD  | 0.89 [0.83, 0.92]         | 0.90 [0.87, 0.93]       | 0.93 [0.92, 0.95] | 1.85 [1.74, 1.96] | 1.02 |
| ASJP40| 0.87 [0.83, 0.92]         | 1.22 [1.15, 1.29]       | 1.30 [1.22, 1.37] | 6.09 [5.44, 6.81] | 1.44 |
| Swadesh100 | 0.85 [0.80, 0.89]     | 0.99 [0.94, 1.03]       | 1.02 [0.98, 1.06] | 2.84 [2.63, 3.06] | 1.24 |
| Swadesh200 | 0.64 [0.46, 0.81]     | 0.89 [0.88, 0.90]       | 0.92 [0.90, 0.94] | 1.95 [1.84, 2.07] | 0.93 |
| Tadmor | 0.82 [0.77, 0.86]         | 1.04 [0.99, 1.08]       | 1.06 [1.01, 1.11] | 3.14 [2.92, 3.39] | 1.23 |

by Swadesh’s original experience-based lists: the 100-item list from Swadesh (1955), which is also used in ASJP (Wichmann et al., 2013), the derived ASJP 40-item list (Holman et al., 2008), and the ABVD list (Greenhill et al., 2008: 210 items). As the only widely used purely data-driven list, we include the Leipzig-Jakarta list (Tadmor, 2009: 100 items).

A convex combination $sc_\lambda := (1 - \lambda) \cdot inf - \lambda \cdot lgc$ of the two partial measures (with $lgc$ inverted because high values are better, and low values are better for $inf$) gives us a single parameter $\lambda$ to optimize towards high ranks for items on the shorter lists. As a quality measure for each possible choice of $\lambda$, we define $m(\lambda, K)$ as the median rank assigned to concepts of a list $K$ divided by the list’s length. A value of 1.0 in this measure thus means that half of the concepts on a list of $k$ concepts end up among the top $k$ positions in our ranking, and lower values imply higher ranks.

The first data column of Table 4 shows the choices of $\lambda$ minimizing the median rank for each list, based on testing all values of $\lambda \in [0, 1]$ with a step size of 0.01. The $\lambda$ which led to the lowest $m(\lambda, K)$ averaged across all five lists was 0.82, which is also very near the optimal value for the Swadesh-100 and Leipzig-Jakarta lists, giving us several reasons to pick this value for our final ranking. To show that the combination of both measures does indeed model the signal in the other lists better than either partial measure could, the last two rows of the table show the $m(\lambda, K)$ values for the two measures individually. It becomes clear that basicness as measured by $inf$ is the better of the two measures, but that refining it by stability as measured by $lgc$ leads to a considerable improvement.
3.2 Ranking correlations

3.2.1 Existing rankings

Having decided on our final measure, we can now compare the resulting ranking with other published rankings. All scoring methods known to us measure stability against replacement (by innovation, semantic change or borrowing). Most do this directly with cognacy percentages, while some do it indirectly by measuring word similarities. Some then find a correlation with an alternative measure. No other automated measure directly attempts to also capture basicness like we do with our definition of average information content.

Dyen et al. (1967) derive glottochronological replacement rates for the Swadesh-200 list from cognacy judgments in 46 pairs of Austronesian languages. Pagel et al. (2007) do the same for (all pairs of) 87 Indo-European languages, and find that frequency in corpora accounts for 50% of the rates’ variation. Petroni and Serva (2011) take the length-normalized Levenshtein distance over word pairs of a concept, and average over all language pairs within one language family, using data from Dyen et al. (1992) and ABVD.11 Peust (2013) represents 13 language (sub-)families by two languages each, and counts for each concept (in a personal list derived from Swadesh’s lists) the number of cognates. Holman et al. (2008) use automatic cognacy judgments from word similarities on ASJP data, and average cognacy percentages over genera as defined in WALS (Dryer and Haspelmath, 2013). The top 40 concepts of this ranking are found to be as good a predictor for the classifications in WALS and Ethnologue (Simons and Fennig, 2018) as the full list, giving rise to the ASJP-40 list.

Rama and Borin (2014) provide an interesting alternative approach by not building on the notion of cognacy. They convert all realizations of ASJP’s concepts within a language family into bags of n-grams, and compute the self-entropy of each bag. This is a simple and indirect measure of the concept’s variability, and inversely, its stability.

Tadmor (2009) uses expert judgments on age, borrowing status, and morphological simplicity, as described in Section 1. He reports scores for the 100 highest-ranking concepts. We refer to the original scores under the name ‘Tadmor,’ and to our re-implementation on current WOLD data (also cf. Section 1) as ‘Tadmor replica.’

---

11 The webpage they link to provided scores for a 100-item subset of Swadesh-200. It has been offline since at least Oct. 2016.
a new approach to concept basicness and stability

Language Dynamics and Change 8 (2018) 157–181

Table 5: Spearman rank correlations with 95% confidence intervals. Overlap is between the respective list and NorthEuraLex.

|            | Overlap | lgc        | inf          | s0.82       |
|------------|---------|------------|--------------|-------------|
| inf        | 0.10 [0.04, 0.16] | 0.43 [0.38, 0.48] | 0.93 [0.92, 0.94] | 0.45 [0.28, 0.60] |
| s0.82      | 0.20 [0.00, 0.38] | 0.35 [0.16, 0.51] | 0.45 [0.28, 0.60] |
| Holman et al. | 99/100 | 0.07 [-0.26, 0.13] | 0.50 [0.45, 0.71] | 0.57 [0.42, 0.69] |
| Rama & Borin | 99/100 | 0.16 [-0.05, 0.35] | 0.30 [0.10, 0.47] | 0.34 [0.15, 0.51] |
| Dyen et al. | 174/196 | 0.31 [0.17, 0.44] | 0.26 [0.11, 0.39] | 0.40 [0.26, 0.51] |
| Petroni & Serva | 93/100 | 0.18 [0.10, 0.25] | 0.54 [0.48, 0.59] | 0.62 [0.50, 0.61] |
| Peust      | 178/185 | 0.35 [0.21, 0.47] | 0.53 [0.21, 0.63] | 0.62 [0.52, 0.70] |
| Tadmor     | 97/100  | -0.01 [-0.21, 0.19] | 0.44 [0.26, 0.59] | 0.40 [0.21, 0.55] |
| Tadmor replica | 626/1498 | 0.18 [0.10, 0.25] | 0.54 [0.48, 0.59] | 0.56 [0.50, 0.61] |
| Pagel et al. | 177/200 | 0.35 [0.21, 0.47] | 0.53 [0.42, 0.63] | 0.62 [0.52, 0.70] |

3.2.2 Results and discussion

Table 5 gives Spearman rank correlations (with confidence intervals) of our measures with the rankings defined by the different published scores, each calculated only on the concepts present in both lists. Our two measures are almost independent, exhibiting only a weak correlation between basicness and stability. As expected, our combined measure is very highly correlated with inf, although lgc also contributes quite a bit of signal, as its correlation with the combined score is significantly higher than that between the two partial measures.

While our combined ranking is clearly correlated with all existing rankings, most strongly with Rama and Borin (2014) and Pagel et al. (2007), and the same pattern repeats itself for the inf measure, only very few lists are clearly correlated with our stability measure. This overall finding is not surprising given that the lgc measure alone was already much worse at sifting out the top concepts during optimization, but the difference provides some interesting insights about the other rankings. The only two other rankings that are correlated with lgc to a relevant degree are Pagel et al. (2007) and Peust (2013), the two approaches that estimate replacement rates using data from some family or predominantly from families also covered by NorthEuraLex. However, lgc is not correlated at all with Rama and Borin (2014) and Tadmor (2009), the two rankings that build on databases with global coverage. This might tell us that purely stability-based rankings built on data from the same regions will tend to agree, but will not generalize to more global data. In contrast, basic-

Downloaded from Brill.com04/26/2019 10:24:06PM via free access
ness seems to be much less data-dependent, and therefore useful for rankings that are more valid cross-linguistically.

To shed further light on why the correlations with the published rankings vary so wildly, we also consider some correlations among published rankings. For this, the three expert-based partial measures of \textsc{wold} are helpful. The borrowing score, representing a major contributor to a concept’s observable replacement rate, turns out to be surprisingly hard to model. Our ranking correlates more strongly than every other ranking with borrowing, although the coefficient is only 0.28 [0.20, 0.35]. At a correlation of 0.14 [0.07, 0.22], the contribution of \textit{lgc} is now more relevant in comparison with \textit{inf}, showing that, as intended, our stability measure is higher for words that are borrowed less.

\cite{Rama2014} turns out to be completely uncorrelated to borrowing at 0.01 [−0.19, 0.20], showing that even global databases will not lead to similar results in the stability ranking; and \cite{Holman2008} actually tends more towards being negatively correlated with borrowing (−0.12 [−0.31, 0.08]). Among the other measures, borrowing is only correlated clearly with \textsc{wold age} (0.16 [0.09, 0.24]) and \cite{Pagel2007} (0.18 [0.04, 0.32]).

\textsc{wold’s age} and \textit{simplicity} scores clearly correlate with all established rankings (albeit with very large confidence intervals due to the differences in list lengths), and of all measures, both correlate strongest with \textit{inf}. Between \textit{inf} and \textit{simplicity}, we measure a high correlation of 0.55 [0.49, 0.60], which is not surprising, as one would expect internally complex words to have higher information content. But the correlation between \textit{inf} and \textsc{age} is also high at 0.49 [0.42, 0.54], which can be interpreted to mean that older words tend to be less complex than newer ones.

There are some interesting patterns not only in the ranking correlations with our measures, but also among the previous rankings we analyzed. For instance, \cite{Rama2014} agrees much more with \cite{Dyen2008} than with \cite{Pagel2007} and \cite{Peust2013}, where the last pair shows a very high agreement at 0.73 [0.65, 0.79]. A systematic analysis of these considerations is beyond the scope of what concerns us here, but we distribute our mapping of the different concept lists as well as the full matrix of resulting ranking correlations as part of the supplementary materials for this article.

4 Evaluating robustness

How certain can we be that a concept that ends up in the top list for any given measure would have ranked equally high on different data? This volatility issue in concept rankings has not been addressed at all by previous work, with the
exception of Pozdniakov (2014), who gives a rough analysis of volatility in a ranking of 100 concepts across many families. In this section, we discuss this issue for our measure from several points of view.

4.1 Resampling language pairs
To evaluate the robustness of our combined measure, we simply use additional language pair samples as we generated them for computing the $lgc$ measure. The results we report in this section are based on a further 1,000 language pair samples. This gives us 1,000 rankings we can summarize to derive two measures of volatility, which provide us with insights about the robustness of our measure against changes in the underlying data.

4.2 Volatility of ranking positions
The lowest mean score across our 1,000 score samples was 1.127 for AND, which is in third position in our ranking based on 100 samples, and the highest mean score was 3.399 for SEPTEMBER, which also comes last in our ranking. The standard deviation of scores averaged across all concepts was 0.090, with the most robust score being that of SALT (1.471 ± 0.055), and the least robust being that of RIPE (2.147 ± 0.127).

A very informative way of looking at the variance of scores is through the window of their implications on ranking volatility. This can be done by considering the range in which each concept falls in 95% of all sampled rankings. For instance, 25 of the 1,000 language chain samples led to ranks above 16 for the concept SALT, and 25 led to ranks below 69. This gives SALT a ranking window size of 53, but puts it among the top-100 concepts on any list derived from our data.

By contrast, the confidence interval for the rank of RIPE is [430, 845], a ranking window of size 415, which could be even larger if the ranking included more than 1016 concepts. Confidence intervals of this size are typical of concepts in the mid-range of the ranking, which means we can expect such positions in the ranking to vary by several hundred positions. Figure 2 visualizes the development of ranking window sizes with ranking positions. The low stability in the mid-range means that small differences in score within that range can be treated as meaningless, and that the ranking should only be seen as providing a classification of a very large number of concepts, extracting basic stable concepts from a large list.

4.3 Certainty of inclusion in concept list
Another perspective on volatility provided by our language pair samples is to ask how certain the ranking values allow us to be about the presence of each
item in a list of top \( X \) concepts. We can frame this question as asking how many of the top \( X \) concepts would have arrived at the top of the list no matter where in the ranking window the score had fallen. The criterion for a concept to be counted as robustly within the top \( X \) concepts, therefore, is whether its ranking position was higher than \( X \) for 95\% of all language chain samples.

For every threshold value \( X \), we can compute the ratio of the top \( X \) concepts that adhere to this robustness criterion. Plotting this ratio for every \( X \), we arrive at the graph in Figure 3. Note that the percentage of certain concepts quickly reaches values close to 50\%, and never drops below 40\% even for values of \( X \) where the effect of exhausting the overall list length does not yet dominate. This means that even though the ranking window sizes may seem rather large, when building a large concept list of, e.g., 500 concepts, our estimates will provide a list of 300 concepts that should be on any such list with a very high degree of certainty, independently of the language pair sample (at least within Northern Eurasia).

5 (Final) ranking and comparison

So far we have introduced new measures of basicness and stability, and determined a convex combination of the two that best emulates the signal contained in a range of existing rankings. We have quantified the overlap with several unordered lists and the correlation with other rankings. This section provides a qualitative comparison. The final ranking (with \( \lambda = 0.82 \)) is given as supplementary material, Figure 4 shows the top 100 concepts.

Even given the volatility inherent in our ranking, it is clear that our ranking confirms most concepts on Swadesh’s lists as well as the ABVD and Leipzig-Jakarta lists. At the same time, it calls into question the inclusion of some
A new approach to concept basicness and stability

Figure 3  Certainty of inclusion (percentages of robust top-X concepts)

concepts occurring on those lists, which receive ranks clearly beyond the top-
200 on any sample of our data. FIFTY [913], DIRTY [861], SICK [854], ROUND [808], TRUE [771], NEAR [706], BLUNT [684], THOUSAND [673], and CORRECT (RIGHT) [647] cannot be recommended for inclusion in a small concept list that is supposed to be used for lexicostatistical research in Northern Eurasia.

Due to the large scope of NorthEuraLex, our ranking also suggests some further items that should be considered for inclusion instead. In addition to some unsurprising candidates like CHEST [75], ARM [54], NEST [70], there are quite a few concepts that were previously dismissed as being too abstract for Swadesh-style lists, with the caveat that some of these may only have ended up in the top region due to peculiarities of the sampled data: PRICE [65], VOICE [93], PLACE (POSITION) [101], DREAM [114], LANGUAGE [128] (cf. TONGUE [102]), THING [161], WORD [167], and TIME [203].

NorthEuraLex already excludes several concepts, such as GUTS and FRUIT, that were found to be too general during data collection. Our ranking shows that the same problem still concerns some concepts on the final list: ANIMAL [821], FLOWER [701], FIGHT [792]; both MAN [364] and WOMAN [418] outrank PERSON [724], SON [44] and DAUGHTER [187] outrank CHILD (YOUNG HUMAN) [343]. Similarly, and due to the lemmas used to look up the concepts, ABVD’s COOK [844] is outranked by BOIL (SOMETHING) [145], BOIL (OF LIQUID) [207] and FRY [386].

Some concepts in ABVD appear to be regional (i.e. less relevant in Northern Eurasia): CLIMB [731], MOSQUITO [769], SPIDER [938], THUNDER [719]. The same applies to Leipzig-Jakarta’s ANT [864]. On the other hand, NorthEuraLex shows its regional bias in natural environment by high positions for WINTER [38] and SUMMER [103], SNOW [59], GRASS [80], FOREST [292], BEAR [283], WOLF [289], and in the mode of subsistence by WOOL [46], COW [108] and MILK [76] (with DRAW MILK [135]), HAY [109], HONEY [109], HORSE [139],
|   | Concept  |   | Concept  |   |
|---|----------|---|----------|---|
| 1 | THAT     | 35 | TOOTH    | 68 | WHITE |
| 2 | DAY (NOT NIGHT) | 36 | LEG      | 69 | SOIL |
| 3 | AND      | 37 | FINGERNAIL OR TOENAIL | 70 | NEST |
| 4 | NOT      | 38 | WINTER   | 71 | END (OF SPACE) |
| 5 | WATER    | 39 | NIGHT    | 72 | DEATH |
| 6 | GIVE     | 40 | NOSE     | 73 | SMOKE (EXHAUST) |
| 7 | THOU     | 41 | STONE    | 74 | FIVE |
| 8 | HEAD     | 42 | DO       | 75 | CHEST |
| 9 | TEN      | 43 | WIND     | 76 | MILK |
|10 | ONE      | 44 | SON      | 77 | LAY (VERB) |
|11 | COME     | 45 | MONTH    | 78 | HOLD |
|12 | HE OR SHE OR IT | 46 | WOOL    | 79 | HORN (ANATOMY) |
|13 | SIX      | 47 | THEY     | 80 | GRASS |
|14 | SALT     | 48 | SEA      | 81 | YOU |
|15 | HUNDRED  | 49 | READ     | 82 | GROW |
|16 | TWO      | 50 | FLY (MOVE THROUGH AIR) | 83 | CARRY |
|17 | TAKE     | 51 | HEAD LOUSE | 84 | BE ALIVE |
|18 | DRINK    | 52 | FISH     | 85 | LONG |
|19 | EYE      | 53 | DIE      | 86 | FATHER |
|20 | THREE    | 54 | ARM      | 87 | HEART |
|21 | WHO      | 55 | MOUNTAIN | 88 | KNOW (SOMETHING) |
|22 | BLOOD    | 56 | DOOR     | 89 | WOOD |
|23 | GO       | 57 | LAKE     | 90 | SUN |
|24 | MOTHER   | 58 | HEAR     | 91 | MELT |
|25 | SEE      | 59 | SNOW     | 92 | OIL (ORGANIC SUBSTANCE) |
|26 | WE       | 60 | HAND     | 93 | VOICE |
|27 | ICE      | 61 | TREE     | 94 | LICK |
|28 | EAR      | 62 | THIS     | 95 | PASS (SOMETHING)* |
|29 | NEW      | 63 | EAT      | 96 | DOG |
|30 | I        | 64 | YEAR     | 97 | SKIN |
|31 | NAME     | 65 | PRICE    | 98 | BOW |
|32 | BURNING  | 66 | BONE     | 99 | FLOW |
|33 | FIRE     | 67 | DRY      | 100 | HAY |
|34 | WHAT     |   |          |   |     |

**Figure 4** Top 100 concepts of our ranking of the NorthEuraLex concept list. Concept names from Concepticon, unless marked with *
HORN (ANATOMY) [79], BOW [98] and ARROW [121], HERD (CATTLE) [270], DRIVE (CATTLE) [216] and HERD [591], BOAT [256] and PADDLE [164], SKI [498] and SLEDGE [713].

Finally, a peculiarity of NorthEuraLex slightly distorting the results is that all concepts are resolved for part-of-speech. This causes both BREATHE [814] and BREATH [722] to be ranked low, because each concept tends to be realized as a derivation from the other in part of the languages.

6 Conclusion

In this paper, we report the first robustness evaluation of a ranking of concepts by adequateness for inclusion in Swadesh-style concept lists. Our investigation reveals a moderate volatility for high-ranked concepts, and a high volatility for mid-ranked concepts in our ranking. A comparison of our ranking to a range of earlier proposals shows that, by adjusting the relative weight of our two partial measures, our ranking can be tuned to agree broadly on a subset of concepts that will receive high scores across datasets and measures, but that the internal order within this top tier of concepts is next to uncorrelated among rankings. The lack of agreement among rankings, as well as our re-derivation of the Leipzig-Jakarta list from more recent WOLD data (Haspelmath and Tadmor, 2009; Tadmor, 2009), indicate that similar amounts of volatility will be present in any ranking. The dependence on data we demonstrated for our ranking has important implications for the design of lexicostatistical resources. If the ASJP-40 and Leipzig-Jakarta lists, which have had a major influence on the concept selection of large-scale lexical databases, are much less well-grounded than commonly assumed, this means we might be faced with a premature thinning of coverage based on unstable results. In future research, we are planning to build on our ranking and the associated volatility information to establish a more well-grounded core for future lists, providing researchers with a pool of hundreds of provably useful candidate concepts, and giving a broader basis to inclusion decisions.

---

12 Pagel et al. (2007) report the uncertainty of their estimated rates as standard deviations, but the robustness against changes in the data is not addressed.
Acknowledgments

This work has been supported by the ERC Advanced Grant 324246 EVOLAEMP —Language Evolution: the Empirical Turn, which is gratefully acknowledged.

References

Bergsland, Knut and Hans Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3(2): 115–153.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097): 957–960.

Buck, Carl Darling. 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages: A Contribution to the History of Ideas*. Chicago: University of Chicago Press.

Dellert, Johannes. 2017. *Information-Theoretic Causal Inference of Lexical Flow*. Ph.D. dissertation, University of Tübingen.

Dellert, Johannes, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2017. NorthEuraLex: A deep-coverage lexical database of Northern Eurasia. http://www.northeuralex.org.

Dryer, Matthew S. and Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures (WALS)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Dyen, Isidore, A.T. James, and J.W.L. Cole. 1967. Language divergence and estimated word retention rate. *Language* 1: 150–171.

Dyen, Isidore, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82(5): iii–132.

Greenhill, Simon J., Robert Blust, and Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4: 271–283.

Haspelmath, Martin and Uri Tadmor (eds.). 2009. *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org/.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42(3–4): 331–354.

Key, Mary Ritchie and Bernard Comrie (eds.). 2015. *IDS*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://ids.clld.org/.
Levenshtein, Vladimir I. 1965. Dvoičnye kody s ispravleniem vypadenij, vstavok i zaměščenij simvolov. Dokladi Akademii Nauk SSSR 163(4): 845–848.

List, Johann-Mattis. 2012. LexStat. Automatic detection of cognates in multilingual wordlists. In Miriam Butt, Jelena Prokić, Thomas Mayer, and Michael Cy Souw (eds.), Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources, 117–125. Stroudsburg, PA: Association for Computational Linguistics (ACL). http://lingpy.org/docu/compare/lexstat.html.

List, Johann-Mattis, Michael Cy Souw, and Robert Forkel (eds.). 2015. Concepticon. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://concepticon.cldl.org/.

Needleman, Saul B. and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3): 443–453.

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

Petroni, Filippo and Maurizio Serva. 2011. Automated word stability and language phylogeny. Journal of Quantitative Linguistics 18(1): 53–62.

Peust, Carsten. 2013. Towards establishing a new basic vocabulary list (Swadesh list). http://www.peust.de/peustBasicVocabularyList.pdf.

Pozdniakov, Konstantin. 2014. O poroge rodstva i indekse stabil'nosti v bazisnoj lek-iske pri massovom sravnenii: Atlantičeskie jazyki [On the threshold of relationship and the “stability index” of basic lexicon in mass comparison: Atlantic languages]. Journal of Language Relationship 11: 187–237. http://jolr.ru/files/(144)jlr2014-2011(187 -237).pdf.

Rama, Taraka and Lars Borin. 2014. N-gram approaches to the historical dynamics of basic vocabulary. Journal of Quantitative Linguistics 21(1): 50–64.

Simons, Gary F. and Charles D. Fennig (eds.). 2018. Ethnologue: Languages of the World, Twenty-first edition. Dallas, TX: SIL International. http://www.ethnologue.com.

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

Tadmor, Uri. 2009. Loanwords in the world’s languages: Findings and results. In Martin Haspelmath and Uri Tadmor (eds.), Loanwords in the World’s Languages. A Comparative Handbook, 55–75. Berlin: Walter de Gruyter.

Wichmann, Søren, André Müller, Annkathrin Wett, Viveka Velupillai, Julia Bischoffberger, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Zarina Molochieva,
Appendix: Information-Weighted Sequence Alignment

In this section, we give a more detailed description of Information-Weighted Sequence Alignment (IWSA), completing the description of how the raw sequence similarity scores \( sc(a, b) \) were calculated for the purposes of this work.

A1 Training the segment distances

The segment similarity scores \( w(x, y) \) for IPA segments \( x \) and \( y \) are PMI scores based on the probability \( p(x, y) \) of \( x \) being aligned with \( y \) in cognate pairs based on counts for a set of cognate candidates, compared to an estimate \( \hat{p}(x, y) \) of that probability on non-cognate words based on 100,000 random form pairs from the NorthEuraLex database:

\[
 w(x, y) := \log \frac{p(x, y)}{\hat{p}(x, y)}
\]

This approach of using alignments of random words to estimate the expected values was also used by List (2012) for estimation of sound correspondences. In our implementation, the \( w(x, y) \) are re-estimated several times based on refinements of the set of cognate candidates through a threshold value on form distances from the previous iteration. In the first iteration, a threshold of 0.35 on the normalized edit distance is used for all form pairs denoting the same meaning. In three further iterations, Needleman-Wunsch distances on the current \( w(x, y) \) are computed by normalization against self-similarity scores as described in the text, and a heuristically determined threshold of 0.6 is used on the resulting form distances to filter the word form pairs with identical meanings for cognacy candidates. Throughout the process, \( \hat{p}(x, y) \) is kept constant, and only the \( p(x, y) \) vary with the current cognate candidate pairs. As in the information content model, additive smoothing distributing 20% of the total number of observations is applied to the correspondence models \( p(x, y) \) and \( \hat{p}(x, y) \) to avoid large values based on little evidence.
A2  Information-Weighted Sequence Alignment (IWSA)

In information-weighted sequence alignment, the dynamic programming table used in our adaptation of the Needleman-Wunsch algorithm is defined by the following recursion (using $\epsilon$ as the gap symbol), which for two phonetic strings $a \in L_a$ of length $m$ and $b \in L_b$ of length $n$ computes the raw sequence similarity score $sc(a, b) := M(m, n)$:

$$
M(\alpha, \alpha) := 0
$$

$$
M(i, \alpha) := M(i - 1, \alpha) + w(a_i, \epsilon) \cdot \hat{I}_{L_a, L_a}^2(a_i, a_i)
$$

$$
M(\alpha, j) := M(\alpha, j - 1) + w(\epsilon, b_j) \cdot \hat{I}_{L_b, L_b}^2(b_j, b_j)
$$

$$
M(i, j) := \min \left( \frac{M(i - 1, j - 1) + w(a_i, b_j) \cdot \hat{I}_{L_a, L_b}^2(a_i, b_j),}{M(i, j - 1) + w(\epsilon, b_j) \cdot \hat{I}_{L_b, L_b}^2(b_j, b_j),} \right)
$$

In these equations, the combined information content $\hat{I}_{L_a, L_b}^2(a_i, b_j)$ is defined as the quadratic mean of both information contents:

$$
\hat{I}_{L_a, L_b}^2(a_i, b_j) := \sqrt{\frac{I_{L_a}(a_i, a_{i-2} \ldots a_{i+2})^2 + I_{L_b}(b_j, b_{j-2} \ldots b_{j+2})^2}{2}}
$$

In the cases of insertion and deletion, this is equivalent to the information content of the segment that was matched to a gap. Otherwise, there is no special treatment of the gap symbol, which is treated like a regular symbol in the global sound correspondence model.