Mathematical supplement (S1)

Johannes Dellert
(johannes.dellert@uni-tuebingen.de)

February 8, 2021

This file is part of the online supplementary materials to:
Johannes Dellert, Niklas Erben Johansson, Johan Frid, and Gerd Carling (2021): *Preferred sound groups of vocal iconicity reflect evolutionary mechanisms of sound stability and first language acquisition: evidence from Eurasia.* In: Antonio Benítez-Burraco and Ljiljana Progovac (eds.): *Reconstructing prehistoric languages*, theme issue of Philosophical Transactions of the Royal Society B.

1 Notation and Input Data

This first section introduces the mathematical structures and the notation which will be used in fully specifying how the sound group stability (SSt) values are derived in Section 2.

1.1 Languages and Families

In our context, a language L can be seen as a collection of consistently segmented IPA sequences representing words from that language. There is some additional structure on the set of languages $\mathcal{L}$, namely a partition into language families $\mathcal{L}_k$ such that $\mathcal{L} = \mathcal{L}_1 \uplus \cdots \uplus \mathcal{L}_n$. The language family assigned to a language L, i.e. the unique $\mathcal{L}_k$ for which $L \in \mathcal{L}_k$, will simply be written $\mathcal{L}(L)$. We do not need any additional structure, such as a phylogenetic tree over the languages.

1.2 Lexical Data and IPA Sequences

Given a concept c from a universal set of concepts C, let $L[c]$ be the set of IPA sequences given by the input lexical database as the possible realizations of the concept c in language L. Such a set can in principle have any cardinality including zero, i.e. there can be concepts for which no forms in language L are given in the database.
All IPA sequences from language sets are implicitly assumed to be segmented consistently according to some common standard. In the case of this paper, the standard we use is determined by previous work done by Erben Johansson et al. [2020]. For a complete specification of this segmentation, it will be necessary to inspect our implementation because we fixed a handful of small problems in the NorthEuraLex 0.9 forms, but a complete list of the segments we use is part of these supplementary materials (S2). The set of all IPA symbols used in a segmented database will be written Σ.

For the purposes of the formalization, it suffices to state that in an IPA string $a \in L_a$ which consists of k segments, we can address the IPA symbol at position i by writing a_i. Symbols can be concatenated by juxtaposition, i.e. the string a can be represented as $a_1 a_2 \dots a_k$. For later notational convenience, we also define a_{-1}, a_0, a_{k+1} and a_{k+2} as a special sentinel symbol # representing the beginning and the end of the word.

1.3 Homologue Annotation

As mentioned in the main paper, we re-use a homologue clustering which was published and evaluated in Dellert [2018]. For the purposes of the formalization, we merely need to state that this clustering partitions the set of words for each concept c into homologue classes of various sizes. We simply number the k_c homologue classes for a concept c, and write them as $H_{c,i}$:

$$\bigcup_{L \in \mathcal{L}} L[c] = H_{c,1} \uplus \dots \uplus H_{c,k_c}$$

To refer to the unique homologue set containing some form $a \in L$, we simply write H_a. Note that identical copies of a can be assigned to different homologue sets (as each language is a collection. not a set, of segmented IPA strings).
Note that we use the neutral term "homologue" here to refer not only to cognacy, but to any type of etymological relationship, including partial cognacy (e.g. derivation from cognate roots) and borrowings. Some of the effects of distinguishing partial cognacy from full cognacy will be achieved by information weighting (see below).

1.4 Alignments

For two segmented IPA sequences $a \in L_a$ and $a \in L_b$, we write the pairwise sequence alignment as $M(a, b)$. This notation is algorithm-neutral, and can refer to the result of different alignment algorithms depending on the context. In every case, an alignment object represents a co-indexing of symbols in both sequences, such that every symbol in either sequences is aligned to either exactly one symbol in the other sequence, or to the gap symbol $-$. The resulting versions of a and b interspersed with gap smybols will be written $M(a, b).a$ and $M(a, b).b$, respectively, so that $M(a, b).a_j$ will be either equal to a_i with some $i \leq j$, or to the gap symbol.

1.5 Sound Groups

The set of IPA segments Σ is divided into overlapping subsets defined by (combinations of) phonetic features. The subsets relevant for this study are named in the main text. In the formalization, they can simply be treated as a list of numbered subsets $\Sigma_1, \dots, \Sigma_m \in \wp(\Sigma)$. Note that in the final analysis, these subsets do not fully cover Σ, as for some sound groups, there were not enough instances of some IPA symbols in the database to make inference possible.

1.6 Concept-Specific Iconicity Data

The values of the Ico (vocal iconicity) variable are previously published scores for some combinations of concepts and sound groups, which will simply be written $Ico(c, \Sigma_i)$ for every relevant concept $c \in C$ and sound group Σ_i.

2 Sound Group Stability (SSt)

Building on the mathematical description of the input data from Section 10, this section formally define how we compute the sound group stability values (SSt) used in our analyses.

2.1 Pairwise Sound Sequence Alignment

The formalization in this section is largely taken directly from Dellert [2018], with some minor changes for notational convenience.

2.1.1 Information Weighting

Writing c_{abc}, c_{abX}, c_{Xbc}, c_{aXc} for the trigram and extended bigram counts extracted from all word forms of a language L, the **information content** of a segment c in its five-segment context $abcde$ (including sentinel symbols) is defined as

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

2.1.2 Information-Weighted Sequence Alignment

The dynamic programming procedure for computing the raw sequence similarity score $sc(a, b) := M(m, n)$ for two IPA strings $a \in L_a$ of length m and $b \in L_b$ of length n is defined by the following recursion:

$$\begin{array}{lcl} M(0,0) & := & 0 \\ M(i,0) & := & M(i-1,0) + w(a_i, \epsilon) \cdot I^2_{L_a,L_a}(a_i, a_i) \\ M(0,j) & := & M(0,j-1) + w(\epsilon, b_j) \cdot I^2_{L_b,L_b}(b_j, b_j) \\ M(i,j) & := & min \left(\begin{array}{c} M(i-1,j-1) + w(a_i, b_j) \cdot I^2_{L_a,L_b}(a_i, b_j), \\ M(i-1,j) + w(a_i, \epsilon) \cdot I^2_{L_a,L_a}(a_i, a_i), \\ M(i,j-1) + w(\epsilon, b_j) \cdot I^2_{L_b,L_b}(b_j, b_j), \end{array} \right) \end{array} \quad (1)$$

For the values $w(a_i, b_j)$, see the next subsection. $I^2_{L_a,L_b}(a_i, b_j)$ is defined as the quadratic mean of information weights assigned to the segments:

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

2.1.3 Estimating Sound Similarity Scores

The similarity scores $w(x, y)$ for IPA segments x and y are pointwise mutual information scores based on the probability $p(x, y)$ of x being aligned with y

in homologue pairs based on counts for a large set of likely homologue pairs, compared to an estimate $\hat{p}(x, y)$ of that probability on non-homologue words:

$$w(x, y) := \log \frac{p(x, y)}{\hat{p}(x, y)} \tag{2}$$

The distribution $\hat{p}(x, y)$ is estimated by randomly sampling as many word pairs (of any meaning) from random language pairs as there are form pairs of identical meaning in the dataset, aligning each pair in the same way that the homologue candidates are aligned, and then counting the number of times each pair of symbols occurred in one column in the resulting alignments. 20% of the overall observation mass is redistributed for Laplace smoothing of the phoneme pair distributions.

The probabilities are not directly based on counts of the number of times each symbol pair was aligned, but each instance in a candidate homologue pair only counts with its combined information content. Using the notation $M(a, b)$ for the optimal information-weighted alignment of a word pair (a, b) and $sc(a, b)$ for the corresponding Needleman-Wunsch score, this way of counting pairs of aligned segments can be written in one expression as follows:

$$c(x, y) := \sum_{L_a, L_b \in \mathcal{L}} \sum_{\substack{c \in C \\ a \in L_a[c],\ b \in L_b[c] \\ sc(a,b) < 1.2}} \sum_{\substack{1 \le i \le M(a,b).len, \\ M(a,b).a_i = x, \\ M(a,b).b_i = y}} I^2_{L_a, L_b}(a_i, b_i) \tag{3}$$

$\hat{p}(x, y)$ is kept constant throughout each iteration of re-estimating $p(x, y)$ from a refined set of cognate candidates. Cognate candidates are selected based on an normalized edit distance threshold (< 0.35) in the initial step, and on a threshold on the Needleman-Wunsch scores (< 1.2) for the current matrix in each subsequent iteration. On our dataset, the values for $p(x, y)$ stabilized after three iterations.

2.2 Estimating Language Pair Divergence

In order to correct for biases caused by the fact that our language pairs represent very different divergence times and therefore expected amounts of change, we measure the average sound stability (i.e the percentage of identical alignments) in homologue pairs from the relevant pair of languages:

$$\bar{s}(L_1, L_2) := \frac{\sum_{a \in L_1,\ b \in L_2,\ H_a = H_b} \sum_{\substack{1 \le i \le M(a,b).len \\ M(a,b).a_i = M(a,b).b_i}} I^2_{L_a, L_b}(a_i, b_i)}{\sum_{a \in L_1,\ b \in L_2,\ H_a = H_b} \sum_{1 \le i \le M(a,b).len} I^2_{L_a, L_b}(a_i, b_i)} \tag{4}$$

The language pair divergence is then simply defined as $d(L_1, L_2) := 1 - \bar{s}(L_1, L_2)$.

2.3 Definition of Stability Percentages and SSt

To measure the cross-linguistic stability for a sound group, we partition all (weighted) instances of sounds from that sound group into four percentages: same-sound stability, shift within group, shift out of group, and gain or loss (i.e. alignment to the gap symbol):

The basic building block for the formal definitions is a sound mapping count *smc* for a segment σ and a set $\Xi \subseteq \Sigma$ which serves as a filter specifying which instances of σ to count:

$$smc(\sigma, \Xi) := \sum_{\substack{c \in C, \\ 1 \leq j \leq k_c}} \sum_{\substack{a,b \in H_{c,j} \\ L_a \neq L_b \\ \mathcal{L}(L_a) = \mathcal{L}(L_b)}} \sum_{\substack{1 \leq i \leq M(a,b).len \\ M(a,b).a_i = \sigma \\ M(a,b).b_i \in \Xi}} I^2_{L_a, L_b}(a_i, b_i) \cdot d(L_a, L_b) \quad (5)$$

As the total count from which the percentages will be derived, we count every alignment including the alignments to the gap symbol (which is not in the set Σ of phonetic symbols): $smc(\sigma, \Sigma \cup \{-\})$.

The same-sound stability *sss* for a sound group Σ_i can now be defined as follows:

$$sss(\Sigma_i) := \frac{\sum_{\sigma \in \Sigma_i} smc(\sigma, \{\sigma\})}{\sum_{\sigma \in \Sigma_i} smc(\sigma, \Sigma \cup \{-\})} \quad (6)$$

Analogously, the shift-within-group percentage *swg* for a sound group Σ_i is

$$swg(\Sigma_i) := \frac{\sum_{\sigma \in \Sigma_i} smc(\sigma, \Sigma_i \backslash \{\sigma\})}{\sum_{\sigma \in \Sigma_i} smc(\sigma, \Sigma \cup \{-\})} \quad (7)$$

The shift-outside group percentage *sog* can be written in the same vein as

$$sog(\Sigma_i) := \frac{\sum_{\sigma \in \Sigma_i} smc(\sigma, \Sigma \backslash \Sigma_i)}{\sum_{\sigma \in \Sigma_i} smc(\sigma, \Sigma \cup \{-\})} \quad (8)$$

Finally, the gain-or-loss percentage *gol* for a sound group Σ_i is

$$gol(\Sigma_i) := \frac{\sum_{\sigma \in \Sigma_i} smc(\sigma, \{-\})}{\sum_{\sigma \in \Sigma_i} smc(\sigma, \Sigma \cup \{-\})} \quad (9)$$

The variable SSt (sound group stability) in our analysis is then simply defined as $sst(\Sigma_i) := sss(\Sigma_i) + swg(\Sigma_i)$.

References

Johannes Dellert. Combining Information-Weighted Sequence Alignment and Sound Correspondence Models for Improved Cognate Detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3123–3133, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

Niklas Erben Johansson, Andrey Anikin, Gerd Carling, and Arthur Holmer. The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology*, 24(2):253 – 310, 2020.