Estimating the areality of phonological segment types using cross-linguistic inventory data

Márton Sóskuthy, *University of British Columbia*

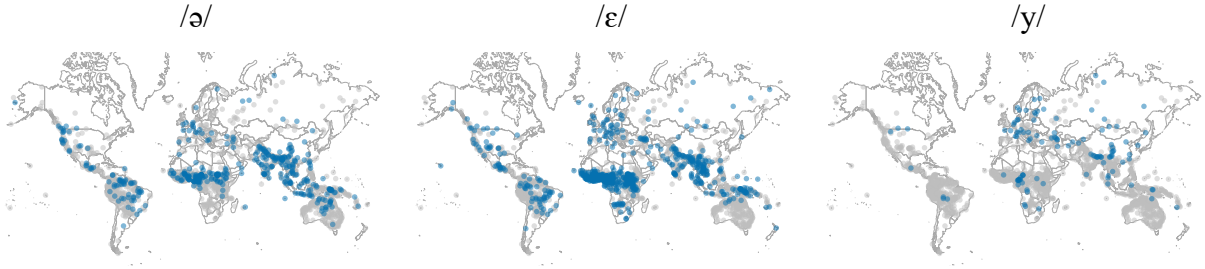
Languages in contact often show convergence, and may come to form *linguistic areas* (Hickey 2017). These areas are recognisable through the higher-than-chance clustering of certain linguistic features. Recent work on areal sound patterns by Blevins (2017) has also suggested that some sound patterns may be more likely to spread than others due to their inherent phonetic saliency. The notion of "higher-than-chance clustering" is key to all the work summarised above, but despite its quantitative ring it is rarely formalised in a rigorous way. We propose a method for quantifying the clustering of contrastive segment types in cross-linguistic inventory data, allowing us to (i) identify segment types that are particularly prone to spreading through contact and (ii) detect linguistic areas defined by these features.

The maps in Fig. A below show the distribution of /ə/, /ɛ/ and /y/ across 2010 languages from the PHOIBLE phoneme inventory database (Moran & McCloy 2019). They clearly illustrate the issues summarised above: are any of these segments more clustered than the others? And where are the clusters? Visual comparisons are difficult as some apparent clusters are simply due to language relatedness, and different segment types can have very different baseline frequencies.

Our key measure *cl* asks the following question: given a language *L* with segment *p*, how likely are its neighbours to also have *p*? This is quantified by looking at the 10 nearest *unrelated* neighbours of *L* and calculating the proportion of *p* among those languages (unrelated neighbours are used to avoid the issue of shared descent). To obtain the clustering of *p* in general, we calculate the average of *cl* across all *L*s with *p* within each language family and average over these averages, obtaining cl_{avg}. cl_{avg} is then compared to a baseline sample representing the null hypothesis that there is no more clustering in *p* than expected by chance. This sample is generated by reshuffling *p* within language families 10,000 times, and calculating cl_{avg} for each reshuffled set. The value of the cumulative distribution function of this reshuffled sample at the original value cl_{avg} is a measure of beyond-chance clustering in *p*, and can also be treated as a one-sided test of significance.

Table B shows these values for the 25 most frequent vowel segment types from PHOIBLE arranged from highest to lowest (some vowels in the middle omitted). /ɛ/ and /y/ show significantly higher-than-chance clustering, while /ə/ does not, consistent with previous reports (see Rolle et al. 2020 for /ɛ/ and Blevins 2017 for /y/). Fig. C shows areal clusters for /ɛ/ and /y/ generated via kernel smoothing from the by-language *cl* values. These clusters are, again, consistent with previous patterns reported in the literature.

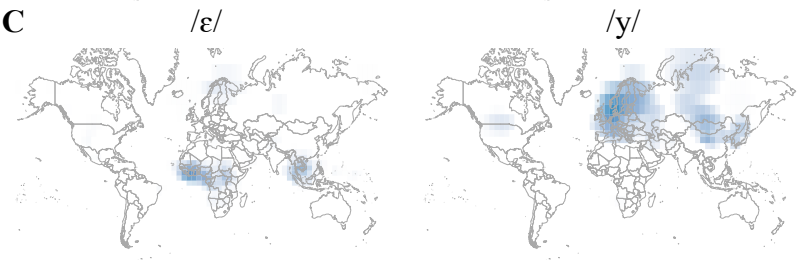
A



B

ɔ	0.9997	ɛ	0.9716	e	0.3745
ɪ	0.9995	y	0.9690	a	0.2277
ʊ	0.9976	au	0.9641	ə	0.1773
o̞	0.9879	ø	0.9575	ʌ	0.1315
ɤ	0.9866	…	…	œ	0.0088

C



References

Blevins, J. (2017). Areal sound patterns: From perceptual magnets to stone soup. Hickey, R. (ed.) *The Cambridge handbook of areal linguistics*, Cambridge: Cambridge University Press, pp. 55-87.

Hickey, R. (2017). Areas, areal features and areality. Hickey, R. (ed.) *The Cambridge handbook of areal linguistics*, Cambridge: Cambridge University Press, pp. 1-15.

Moran, S. & McCloy, D. (2019). PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History. Available online at http://phoible.org, Accessed on 2022-02-21.

Rolle, N., Lionnet, F. & Faytak, M. (2020). Areal patterns in the vowel systems of the Macro-Sudan Belt . *Linguistic Typology*, *24*(1), 113-179.