The half-way similarity avoidance rule replicated using phonetic data from European language varieties

Matías Guzmán Naranjo, University of Tübingen, mguzmann89@gmail.com
Søren Wichmann, University of Kiel

Short title: The half-way similarity avoidance rule

Acknowledgements

SW was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2150 – 390870439 (ROOTS) and grant no. G2021125001L of the International Collaboration Program of Nankai University. MG was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 834050), within the Crosslingference project; and by the Emmy-Noether project 'Bayesian modeling of spatial typology' (grant number 504155622).

Competing Interests:

The author(s) declare none.

The half-way similarity avoidance rule replicated using phonetic data from European language varieties

Abstract: Previous work using lexical data from around the world has suggested that distances among language varieties are distributed such that varieties are typically either rather similar, qualifying as dialects of one another, or rather dissimilar, qualifying as different languages, with a scarcity of varieties that are around half-way similar. Wichmann (2019) observed that there is a bimodal distribution of distances with two roughly normal distributions separated by a valley. The previous work was based on a database mostly containing either descriptions of single languages or surveys covering several close varieties, so the bimodal distribution could potentially be an artifact of the underlying sample. Here we test whether a similar distribution is found when using another source of data and an unbiased sample drawn from the cells of a geographical grid (of Central Europe). The data consists of 18 lexemes from 274 doculects. Using Bayesian Beta regression and leave-one-out cross-validation, we show that the data follows a bimodal distribution which is robust to sampling, and also to at least some aspects of the data (coarse- vs. fine-grained phonetic transcriptions).

Keywords: languages vs. dialects, phonetic distance, European languages, dialectometry, language classification

1. Introduction

In the past, attempts to distinguish languages and dialects have tended to either be qualitative in nature, subject to discussion and negotiation, sometimes with political implications (Van Rooy 2020), or else they have relied on some cut-off of intelligibility among speakers or along a measured continuum of similarity (typically lexical similarity) pertaining to languages (Voegelin and Harris 1951). Thus, attempts to distinguish dialects from languages

have begged the question whether the distinction is a real one, inherent to languages, or whether it merely reflects a desire to impose such a distinction. Such a desire is completely understandable and reasonable, because we need to distinguish languages for many purposes. These include both scientific purposes, such as the organization of knowledge (catalogs, indexes, maps) or the analysis of linguistic diversity, and more applied purposes, such as revitalization strategies, literacy campaigns, translation efforts, and so on. Strong as this desire may be, it does not constitute evidence that the object of the desire actually exists independently of it.

In Wichmann (2019) it was observed that distances among varieties of related languages are distributed in such a way that they will tend to be either rather similar or rather dissimilar, with a dearth of intermediate cases. It seemed that a real difference between what we might call dialects and what we might call languages had been found. The paper drew upon word lists for 451 doculects from 15 groups of related languages in the ASJP database (Wichmann et al. 2018). It was shown that the differences between the word lists, when displayed in density plots (essentially equivalent to smoothened histograms), tended to show bimodal distributions (double humps reminiscent of a dromedary camel). The ‘valley’ or near-gap in the distributions occurred around a similarity of 0.51 on a scale from 0 to 1, using the normalized Levenshtein distance (for the definition of which look further on in the present paper). Using a dating technique from Holman et al. (2011), this similarity could be translated into a time separation of around 1350 ± 300 years. Not only did this cut-off emerge through an objective procedure, it also yielded distinctions which could be described by the dialect vs. language labels as they might plausibly be applied.

A potential issue with Wichmann (2019) is that it relied on samples of language varieties that could have introduced a bias. The ASJP database contains data from many linguistic surveys, and these tend to be directed at the documentation of what is perceived to be ‘dialects’ of some ‘language’. Often surveys focus on a particular geographical area. Survey data, however, is only included on an opportunistic basis, whereas the main aim of the database is to cover as many different ISO-639-3 entities (‘languages’) as possible. So, the database mainly contains ‘language’ data, with some additional ‘dialect’ data, thus perhaps inadvertently being biased towards a ‘language/dialect’ distinction. Some resampling was carried out towards an attempt to control for such a bias, but the limitations of this attempt were acknowledged.

Here we aim at a replication of the general findings of Wichmann (2019), using a different type of initial, unbiased sample, albeit one which is smaller. A survey aimed at simply covering whatever variety is spoken in the cells of a geographic grid should be unbiased, whereas a survey aimed at 'dialects' (close varieties) of a 'language' might miss intermediate varieties (less close varieties) between 'languages', something which might be responsible for the valleys in the density plots seen in the earlier study. The European part of the database described in the next section allows us to work with a grid, removing biases from the sample.

Introducing a grid-based sample is not the only motivation for moving to data other than those employed by Wichmann (2019)—the ASJP database would also allow for drawing such a sample.[1] Additionally, we want to see whether the observations of Wichmann (2019) generalize using data that are very different in nature. The ASJP data comes in transcriptions that merge many phonological distinctions, whereas the data we will employ here comes in maximally fine-grained phonetic transcriptions.

2. Materials and methods

2.1 *Dataset*

We use the Sound Comparisons database (Heggarty et al. 2019), which contains around 50,000 narrow transcriptions of words for dozens of cognates in over 600 language varieties[2] around the world.[3]

Since we want to reduce sample biases by controlling for geographical coverage, we are interested in being able to sample from a grid which is as fine-grained as possible, while, at the same time, having as few empty cells as possible. To illustrate why this is important, consider the following example of the nature of a sample not having this kind of geographical control. The ASJP database (Wichmann et al. 2018) includes 47 word lists from a linguistic survey of northern Pakistan (Bakstrom and Radloff 1992). Of these, 26 represent varieties of

1 ASJP has a good geographical coverage. But it needs to be mentioned that metadata relating to locations are inconsistent and therefore not very adequate. Most often, varieties of a language as defined by an ISO 639-3 code are assigned the same geographical coordinates, but sometimes varieties are assigned different locations.

2 After post processing we ended up with 247 varieties.

3 The data is freely available from https://soundcomparisons.com/. To download it, the reader has to select the Europe dataset and, for each cognate set, download the csv file. The process needs to be repeated for the individual datasets (Germanic, Celtic, Slavic, Romance). Along with this paper we only share the processed data.

Shina [scl]. Thus, for no particular reason other than the availability of Bakstrom and Radloff (1992), these data on Shina dialects would play an important role in a study of how to distinguish languages and dialects drawing upon a sample taking from ASJP and defined genealogically such as to include Indo-Aryan languages or defined geographically such as to include northern Pakistan. Sampling from a geographical grid, however, would remove the bias, but only if cells are filled at least to a reasonable extent. Having a grid-based sample with a lot of empty cells obviously defies the purpose of a grid-based sample. For instance, we may end up mostly having language varieties from northern Pakistan in our sample if, say, we sampled from an area including Pakistan, but (1) were largely limited to data from Bakstrom and Radloff (1992) and (2) allowed for a lot of empty cells.

The Sound Comparisons database lends itself to a grid-based approach. According to Heggarty et al. (2019: 281), "[s]ampling of language varieties has been determined by two main criteria: to be representative of linguistic and dialectal diversity, and urgency in the face of the imminent extinction that hangs over much of that diversity. The complex balance between these criteria often overrides the default of sampling evenly through geographical space." Even if, as is admitted, sampling is not geographically even, we can control for this by sampling from the sample. Since Europe is the most densely covered area in the database, we have selected an area within Europe. Figure 1 shows the locations of all the European language varieties of the database. Further on (section 2.4 and Figure 2) we describe how we extract a Central European grid from the area for the purpose of the present investigation.

The data across language varieties need to be comparable, so from the word lists we select only 18 items, namely the items that are attested across all the Celtic, Germanic, Romance, and Slavic languages. The items in question are: 'eight', 'five', 'four', 'full', 'grain', 'hundred', 'name', 'nine', 'one', 'salt', 'seven', 'six', 'ten', 'three', 'tongue', 'two', 'wind', and 'young'.[4] If we were to measure the similarity of different varieties using different numbers of cognates by availability, we would be introducing a confounding factor. While only 18 cognates may sound like too little, Section 3.2 presents some robustness checks showing that our results are likely not due to having too few cognates.

4 The reason why there are mostly numerals in the dataset is that the Sound Comparisons dataset is based on independent lists of cognates for each Indo-European subgroup. Numerals seem to be stable across all of the four included subgroups.

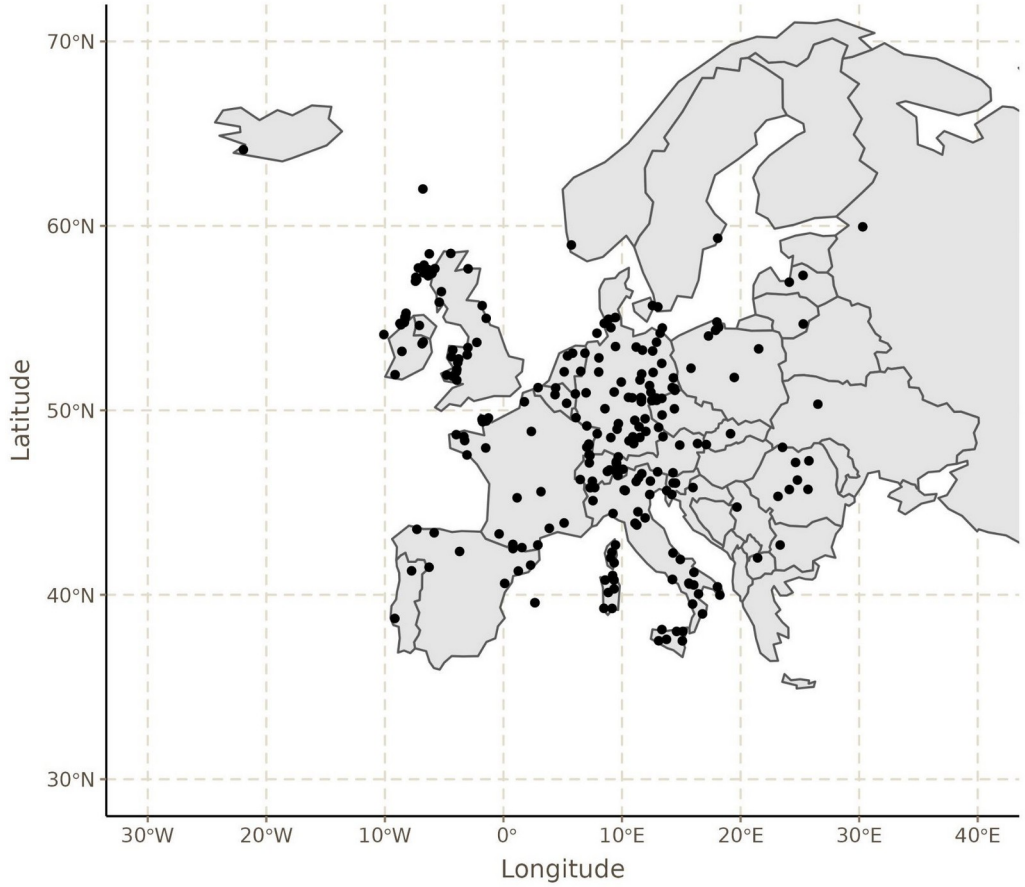

Fig. 1. Location of language varieties represented in the Sound Comparisons database.

Likely due to the fact that several transcribers contributed to Sound Comparisons, the phonetic transcriptions contain some inconsistencies. These concern the use of different diacritics to mark the same phonetic process, different unicode characters for what is arguably the same phone or diacritic, and the use of some non-standard diacritics. We did our best to normalize the transcriptions by (1) unifying different UTF-8 characters, (2) using the same diacritic systematically, and (3) removing diacritics used for features that were not marked consistently.[5]

The types of changes we made to phonetic strings include the following:[6]

- We removed all non-IPA-like symbols like: []()+_ , etc.

5 For example, stress was only marked for a handful of dialects.
6 The full list of changes can be found in the supplementary materials. These are available at https://doi.org/10.5281/zenodo.7752997 .

- We transformed single unicode characters for symbols like <ä> to combinations like <a> + <¨>, which are visually indistinguishable, but different for the computer. The reason is that some annotators used the single symbol, while others had picked the two-symbol combination.
- We removed stress marks on vowels (e.g. <í>) because it is unclear what they mean, and these are non-systematic throughout the dataset
- We converted non-standard symbols like <ʔ> (U+0241) or <g> (U+67) to their closest IPA symbol: <ʔ> (U+0294) and <g> (U+0261).

Some caveats are in order. Certain symbols are normally not used in IPA, and were only present in a handful of transcriptions, like φ and superscript ˢ. It is difficult to know exactly what the transcriber meant by these, but we tried to approximate their most likely equivalent (in these cases ɸ and ʃ). Other cases were even more ambiguous, like the use of ʂ, which normally represents s^w in African linguistics, but which we think was meant to represent ʂ. However, cases like these were very rare, with only a dozen occurrences, and should not make a noticeable impact on our dataset.[7]

This process left us with a total of 1529 different symbols. The reason for this very large number is twofold. First, we are counting long and short versions of the same phoneme as different symbols. Secondly, the transcriptions vary in degree of precision, with some transcriptions using three or four diacritics on the same symbol.

One additional step taken was to randomly select only one transcription per dialect. In a small number of cases, the dataset records several alternative cognates for a variety.[8] Since this is not consistent, and only happens sporadically, we decided to keep only one form per variety.

In order to be able to calculate phonetic distances between words and thus dialects, we built a feature matrix including all symbols in our dataset. We use the Panphon phonological feature matrix as a starting point (Mortensen et al. 2016) and added several additional features to be able to distinguish all contrasts marked in the dataset (extra short, half long, centralized, retracted, advanced, lowered, raised, non-audible and mid-central). While, strictly speaking, we are not capturing phonological contrasts but rather phonetic

7 See the supplementary materials for the full list of changes.
8 A couple of entries also list non-cognates. We removed these.

ones, we use an approach based on phonological-like features to capture similarities between sounds.

2.2 *Distance metrics*

Distances between pairs of lexical items form the basis of our computation of phonetic distances between language varieties in the dataset. In order to measure the phonetic distance between strings, we need to calculate a phoneme substitution cost corresponding to an assumed distance between two different phonemes. In the literature, there are two ways of doing this. The simplest approach is to assume a uniform cost of 1, independently of which segments we are comparing. For example, the substitution cost of [t], [d] and [x] would be 1 for any transition among the three different sounds. Alternatively, we could claim that [t] is more similar to [d] than it is to [x], which should incur different substitution costs, or that vowels are more similar to each other than they are to consonants because they have more features in common, and that their substitution costs should thus reflect this difference in similarity. There are multiple possibilities for measuring phonological distances of phonemes (Beniamine 2017, Beniamine and Guzmán Naranjo 2021), as well as various implementations (see Mortensen et al. 2016, Dellert and Jäger 2017, List 2017, Kilani 2020), and as far as we are aware, there does not seem to be an agreed-upon choice in the literature. In this paper we used a binary distance metric based on a feature description.[9]

Given a phoneme substitution cost (whether 1 or other), we can use the normalized Levenshtein distance to calculate the distance between translational equivalents in different varieties. The Levenshtein distance (also called edit distance) counts the number of operations required to transform string *s* into string *t*. The possible operations are deletion, insertion and substitution, and each one has a cost assigned to them. We assign a cost of 1 to insertion and deletion, and use the phoneme substitution costs for substitution. The normalization, which is quite commonly used, consists in dividing the distance by the length of the longest word, resulting in a number between 0 and 1.

For example, the word for *five* is recorded in Stavanger as [fɛm] and in Swedish as [fɛm̃]. The raw edit distance between the two words would be 1 if we counted [ɛ] and [ɛ̃] as completely different segments and assign a substitution cost of 1. However, we have a

9 One potential downside of using a binary distance metric is that it cannot distinguish between features which are *off* and features which do not apply to a given phoneme usually marked as 0 in a phonological feature matrix.

substitution cost of 0.17 between nasal and non-nasal vowels, meaning that the edit distance is 0.17, which we divide by 3 (the maximum length of either string) to get 0.06 (rounding the value).

Recall that our data consists of phonetically transcribed realizations for a series of words in many variants. For example, we have the realization of words like *four* and *five* in Faroese ([ˈfʊ̈ɪ̯ä], [fɪmˑ]), Stavanger ([ˈfiːʁə], [fɛm]), Swedish ([ˈfyːɾä], [fɛ̃m]), etc. For each word we can calculate the distance to its translational equivalent in another language variety, as shown in Tables 1-2. Table 1 shows the Levenshtein distance using no phoneme substitution costs (all substitution costs = 1), and Table shows the distance using costs.

Table 1. Distances among selected words in three Scandinavian language varieties using Levenshtein Distance without phoneme substitution costs

	four			*five*		
	Faroese [ˈfʊ̈ɪ̯ä]	Stavanger [ˈfiːʁə]	Swedish [ˈfyːɾä]	Faroese [fɪmˑ]	Stavanger [fɛm]	Swedish [fɛ̃m]
Faroese	0	0.8	0.6	0	0.67	0.67
Stavanger	0.8	0	0.75	0.67	0	0.33
Swedish	0.6	0.75	0	0.67	0.33	0

Table 2. Distances among selected words in three Scandinavian language varieties using Levenshtein Distance with segment substitution costs

	four			*five*		
	Faroese [ˈfʊ̈ɪ̯ä]	Stavanger [ˈfiːʁə]	Swedish [ˈfyːɾä]	Faroese [fɪmˑ]	Stavanger [fɛm]	Swedish [fɛm̃]
Faroese	0	0.48	0.39	0	0.1	0.14
Stavanger	0.48	0	0.35	0.1	0	0.06
Swedish	0.39	0.35	0	0.14	0.06	0

Given the kind of information on word distances illustrated in Table 2, there are several alternatives for calculating the overall distance between languages. The most straightforward is to take the mean distance across all words for each pair of languages. For the example in Table 2, the overall distance between Faroese and Stavanger would be 0.29, between Faroese and Swedish 0.265, and the distance between Swedish and Stavanger would be 0.205. If we assume substitution costs of 1 for all phonemes, as in Table 1, the process would be analogous.

In this paper we explore both the approach using uniform substitution costs and the one using feature-based substitution costs. In the following section we also compare our method to two additional methods described in the literature.

2.3 *Validation of distances*

Before employing the distances in further analyses it should be decided which measure more adequately represents distances among speech varieties. We employ two strategies for comparing the performance of the distance measures.

The first evaluation strategy is to compare linguistic distances to geographical distances. It is a standard expectation in dialectometry since the work of Séguy (1971) that geographical distances among language varieties are correlated with their linguistic distances (see further references in Holman et al. 2007:395). So, among alternative linguistic distances, the one that leads to a better congruence with geography would appear to be more adequate. Here we use the Great Circle Distance (GCD) and coordinates as given in the Sound Comparisons database. Using the GCD rather than some more sophisticated measure of travel

distance is justified by a study that introduces such a measure, but nevertheless concludes that "for a correlational study limited to distances up to 2000 km the GCD will do" (Wichmann and Hammarström 2020:5). The vast majority of distances computed in the present study are smaller than 2000 km , and none exceeds 2600 km. Linguistic and geographical distance matrices are compared through CADM (congruence among distance matrices), as implemented in the CADM.global() function of the R package ape (Paradis and Schliep 2019; see Legendre and Lapointe 2004 for an introduction to CADM). We report on Kendall's coefficient (W).

The second evaluation strategy consists in comparing trees based on different linguistic distances with some yardstick to see which tree is more similar to the yardstick. The yardstick used here is the classification of the lects according to Glottolog (Hammarström et al. 2021). Creating a Glottolog tree for the Sound Comparisons doculects was done as follows. First we assigned ISO 639-3 codes to each lect. In this step we let ourselves be guided by links to Glottolog and/or Wikipedia that are given on the page for nearly every lect in the online Sound Comparisons database.[10] Next, we prepared a simple text file with one line per lect, containing its name and its Glottolog classification as expressed by decreasingly inclusive groups from the family level to the ISO 639-3 code level. The file was submitted to Greenhill's treemaker tool (Greenhill 2018) in order to produce a tree in newick format. Using a variety of tree comparison methods implemented in the R libraries TreeDist (Smith 2020) and Quartet (Smith 2019), this could then be compared to trees based on our weighted and unweighted Levenshtein distances (henceforth LD-W and LD-UW) and produced in MEGA (Kumar et al. 2018) using the popular Neighbor-Joining algorithm (Saitou and Nei 1987).

Two additional sets of distances were tested. The first was computed by first transforming the phonetic transcriptions into ASJPcode (Brown et al. 2013) and then performing an unweighted Levenshtein distance. The second approach is to use LingPy's (List 2022) distance metric. We use the pw_align() function, with global alignments, to produce a phonological distance, using default settings. This function implements the distance calculation proposed by Dawney et al. (2008). The results of comparing linguistic and geographical distances are shown in Table 3. They show congruences which are better

10 For 4 lects we could not assign ISO 639-3 codes. They are listed as follows in the source: 'Italy: N. in S.: Picerno', 'Italy: N. in S.: Tito', 'Italy: N. in S.: San Fratello', 'Italy: N. in S.: Novara'. Picerno and Tito are Gallo-Italic varieties of the Basilicata region of Southern Italy, and San Fratello and Novara are Gallo-Italic varieties of Sicily. Neither the Basilicata nor the Sicilian varieties seem to be taken into account in Glottolog.

for the weighted distance, but not by much. Overall, the ASJPcode and Lingpy produce slightly worse fits with the geographic distance.

Table 3. Kendall's W between linguistic and geographical distances for all distance matrices. Best score in bold.

distance	LD-W	LD-UW	Lingpy	ASJP
Kendall W	**0.70**	0.69	0.68	0.66

Results of two tree comparison metrics are shown in Table 4. The Steel-Penny variant of the quartet distance is a normalized quartet distance computed as the quartet distance divided by the sum of the quartet distance and the quartet similarity (or total number of quartets). It runs in the [0-1] range. The Robinson-Foulds distance counts the absolute number of nodes differing between two trees. The quartet distance points to LD-UW as performing best, but with LD-W as a very close contender. Because of noise due to sparseness of data and a reference tree which is not fully resolved, we do not invest confidence in the ability of the test to reliably distinguish LD-UW and LD-W. It does seem clear, however, that Lingpy and ASJPcode transcriptions perform less well. As for the Robinson-Foulds distance, which is less fine-grained than the quartet distance, the result is a tie throughout. In sum, it is hard to exclude either LD-W or LD-UW from further analysis, whereas it seems safer to exclude Lingpy and ASJP.

Table 4. Distances between Neighbor-Joining trees using LD-W vs. LD-UW according to two tree comparison metrics. Best scores in bold.

LD-W	LD-UW	Lingpy	ASJP	R package	metric
0.163	**0.162**	0.172	0.168	Quartet (Smith 2019, Sand et al. 2014)	Steel & Penny (1993)
250	250	250	250	TreeDist (Smith 2020)	Robinson and Foulds (1981)

The two Neighbor-Joining trees and the Glottolog-based tree are supplied as Supplementary Materials.

2.4 *Grid sampling*

In order to determine the design of a grid we were interested in (1) maximizing the area covered as well as (2) the number of grid cells (i.e. minimizing their size), while (3) minimizing the number of empty cells. Further specifications of these criteria and their respective rankings would be needed for a unique solution to the optimization problem, but such specifications would be arbitrary, so we settled on an ad hoc, experimental solution, shown in Figure 2. The grid consists in 45 cells of size 3°×3°, of which just six are empty.

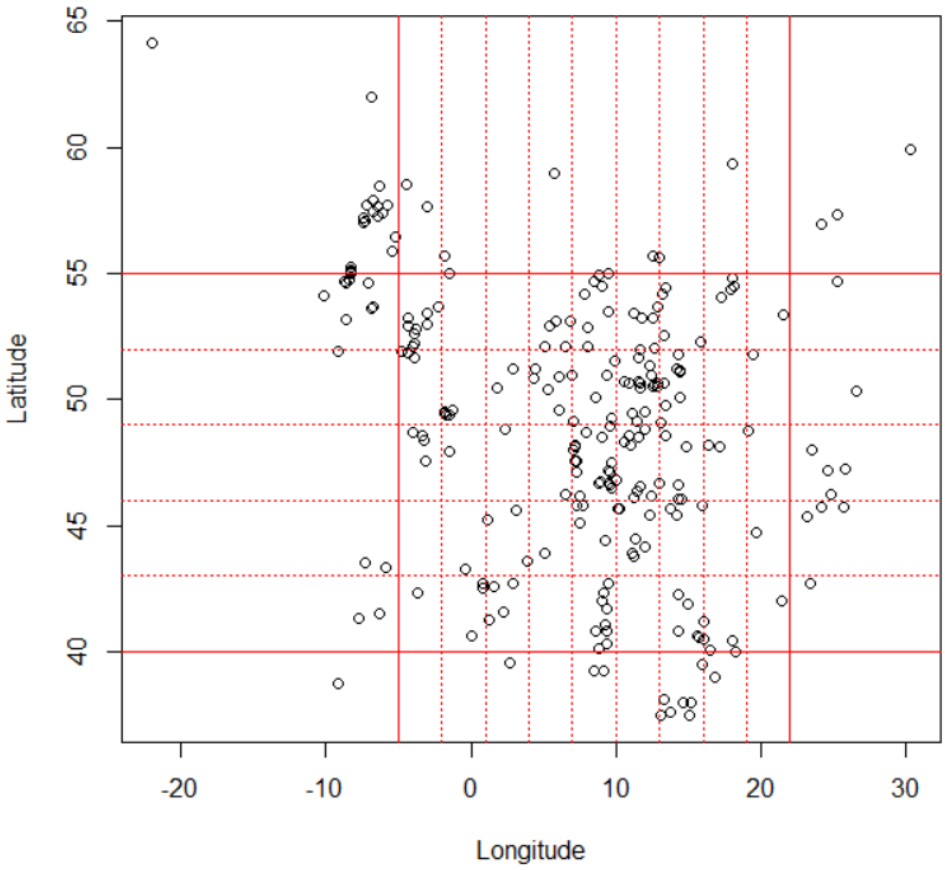

Fig. 2. Geographical grid used for sampling. Solid lines show the bounds and dotted lines the grid cells.

Given the grid in Figure 2, we can produce a balanced sample of language varieties by choosing one random location per grid cell. This sampling procedure may be repeated with replacement several times over in order to capture the variation in the data.

3. Results

3.1 *Density plots*

Figures 3 and 4 show density plots of distances among language varieties using the grid sampling method described in the 'Materials and methods' section. Figure 3 represents distances based on phonological substitution costs and Figure 4 distances based on uniform substitution costs.

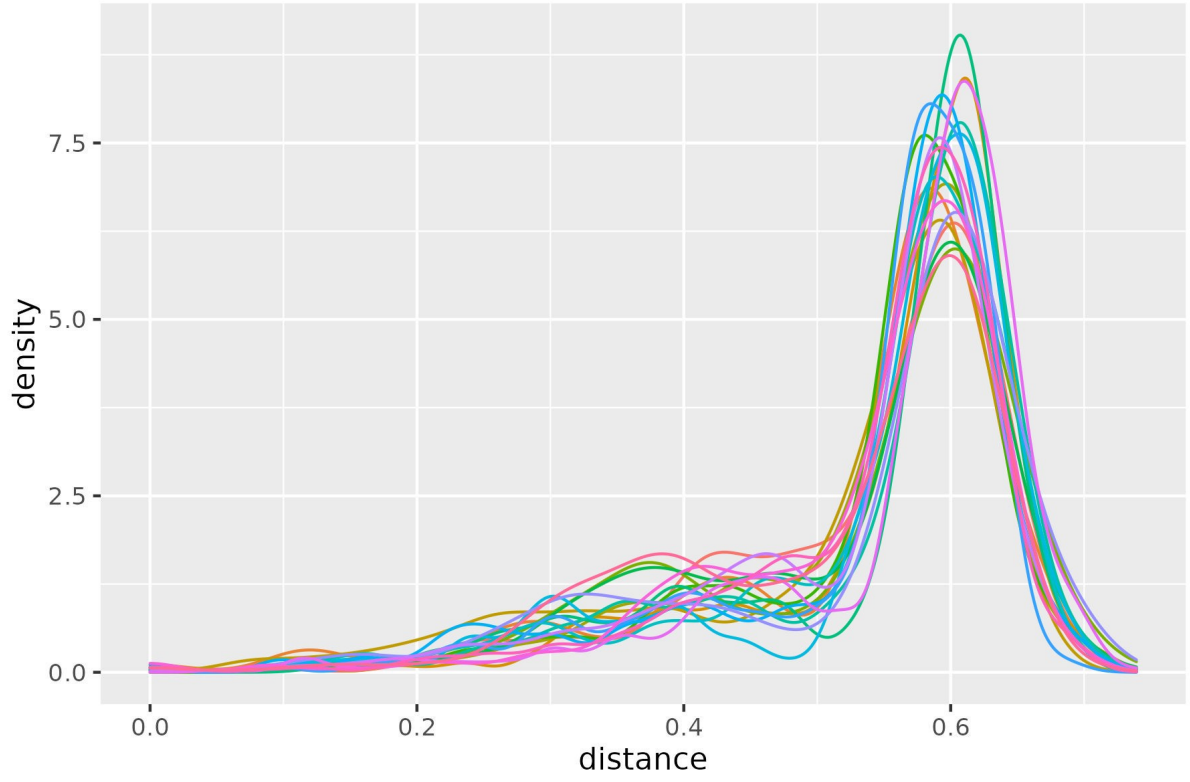

Figure 3. Density plot of pairwise distances for 20 grid samples of language varieties using the normalized Levenshtein distance with substitution costs.

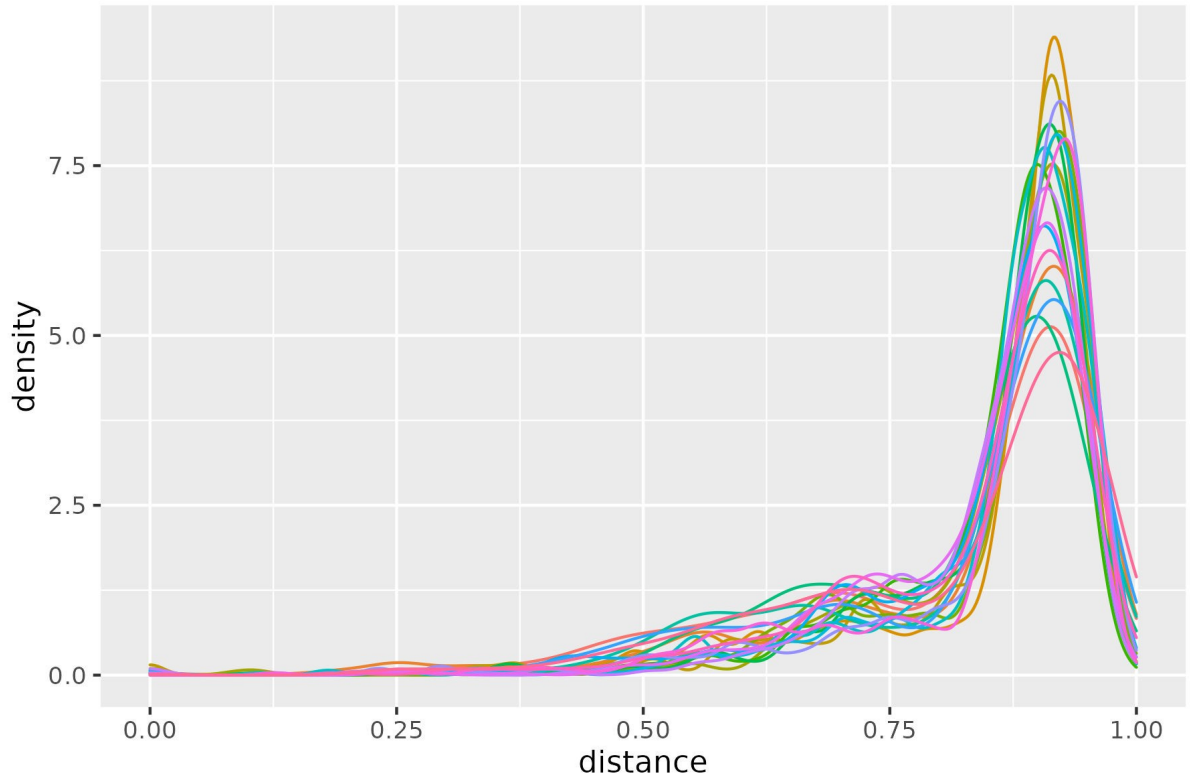

Figure 4. Density plot of pairwise distances for 20 grid samples of language varieties using the normalized Levenshtein distance without substitution costs.

At first sight, it is not obvious that there is a valley separating two distributions in either plot. What we see is a large peak at 0.6 in Figure 3 and at 0.9 in Figure 4, both preceded by a relatively flat section. However, we do observe small dips at around 0.5 in Figure 3 and around 0.8 in Figure 4.

We can nonetheless ask whether these distributions reflect an underlying single distribution or a mixture of two distributions. To address this question we fitted two Bayesian models using Stan (Carpenter et al. 2017) and BRMS (Bürkner 2017) on one of the samples from the dataset (as described above). One model (M1) is composed of one Beta family, and another model (M2) has a mixture of two Beta families for each distribution. Importantly, we could not fit a model with a mixture of three families for either distribution. The models failed to converge and there were label-switching issues.

We want to compare whether a single beta family or a mixture of two beta families captures the data better. We first do a visual inspection using posterior predictive checks on both models (Gabry et al. 2019). This consists of generating random data from the posterior

distributions obtained in the model fit, and comparing the density distribution of the obtained random data to the density distribution of the original data. If a model fits the original data well, the generated data should have a density distribution similar to the original one. Figures 5 and 6 show the posterior predictive checks for M1 and M2, respectively, for the distance using phoneme substitution costs.

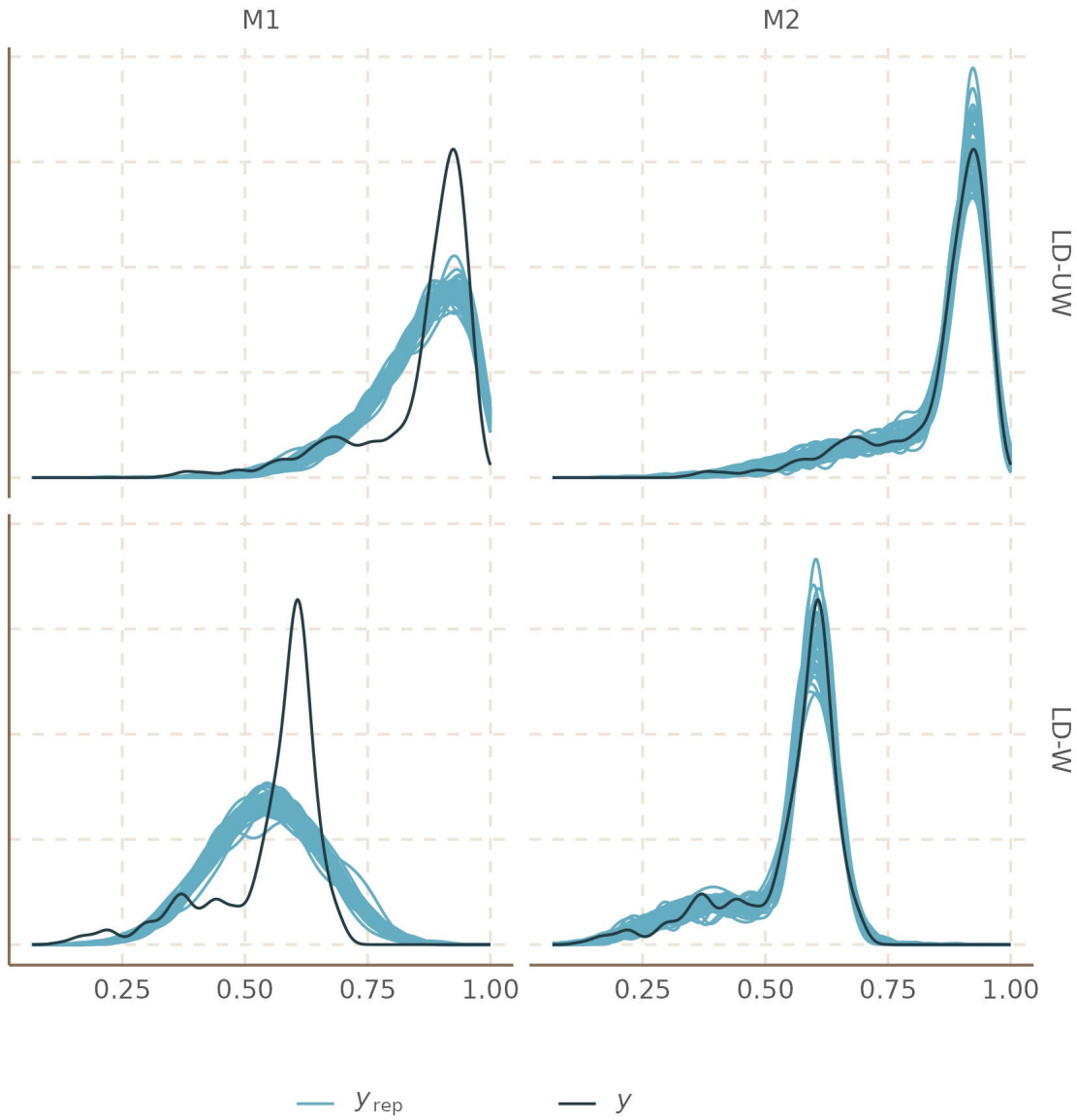

Figure 5. Posterior predictive checks for models M1 and M2 with phoneme substitution costs (LD-W) and without phoneme substitution cost (LD-UW). Y is the density of the sample data used on the model, Yrep is the density of the predicted data from the model.

In Figure 5, the dark blue line shows the density distribution of the data sample used to fit the models, while the multiple light blue lines show the random samples from the posterior distribution. In this case it is clear that with both distance metrics, the M2 model with a mixture of two Beta families fit the data much better than the model with a single Beta family.

The second method we explore in order to evaluate model fit is (approximate) leave-one-out (LOO) cross-validation (Vehtari et al. 2017). In LOO[11] we remove one observation from the dataset, fit the model, try to predict that left-out observation, and then evaluate how successful the model was in making the right prediction. We repeat this process for all observations. Due to computational constraints, we employ an efficient approximation using Pareto-smoothed importance sampling (see Vehtari et al. 2017 for details). To evaluate model performance, we use the expected log predictive density (ELPD), which captures how much probability density a model assigns to the real value of the predicted observation.[12]

Absolute ELPD values are not very informative in our case, but the difference in ELPD between two models is. Tables 5 and 6 show the ELPD difference between the best model (M2) and the second-best model (M1), as well as the Standard Error of the difference. The best model is assigned values of 0, serving as a baseline. In both cases, the model using a mixture of two distributions (M2) performed considerably better than the model with a single distribution. Interestingly, the difference is much clearer when we take substitution costs into consideration.

Table 5. ELPD difference between M1 and M2 on a grid-sample of the Levenshtein distances with variable phoneme substitution costs.

	ELDP difference	SE difference
M2	0	0
M1	-271	21.7

11 This approach to model evaluation is an alternative to the more traditional Bayes Factor, and it is similar in principle to (W)AIC. The drawback of Bayes Factor is that it can be extremely sensitive to prior choice, and in some cases overconfident with misspecified models (Oelrich et al. 2020). In comparison to WAIC, LOO has been shown to be less error-prone, and more accurate overall (Vehtari et al. 2017).

12 The ELPD can be thought of as a version of MSE or MRSE, but it takes into account the posterior uncertainty in the predictions.

Table 6. ELPD difference between M1 and M2 on a grid-sample of the Levenshtein distances with uniform costs.

	ELPD difference	SE difference
M2	0	0
M1	-134.2	15

Tables 5 and 6 clearly show that M2 is much better than M1 in both cases, with and without substitution costs, although the difference is larger if we take substitution costs into consideration. This result suggests that using variable phoneme substitution costs makes the bimodal nature of the distribution more salient than when no (or uniform) costs are used.

Because we are estimating the parameters of the distributions, we can also try to visualize what the underlying distributions would be, if they were independent. For reasons of space we only show this for the distances using phoneme substitution costs, but a similar behavior is found for the regular Levenshtein distances. Figure 6 shows posterior samples from these estimated distributions, as well as the observed grid-sample used to fit the model M2. We can see that the distribution for shorter distances (or the distribution of dialects) is much wider than the distribution for longer distances, which is concentrated at 0.6 and has very little spread. We also observe that while there is some overlap (the mean intersection of the posteriors is at 0.536, with the whole posterior lying between 0.527 and 0.547, and the mean posterior overlap of the distributions is 0.07), a large portion of the distributions does not overlap and is clearly differentiated.

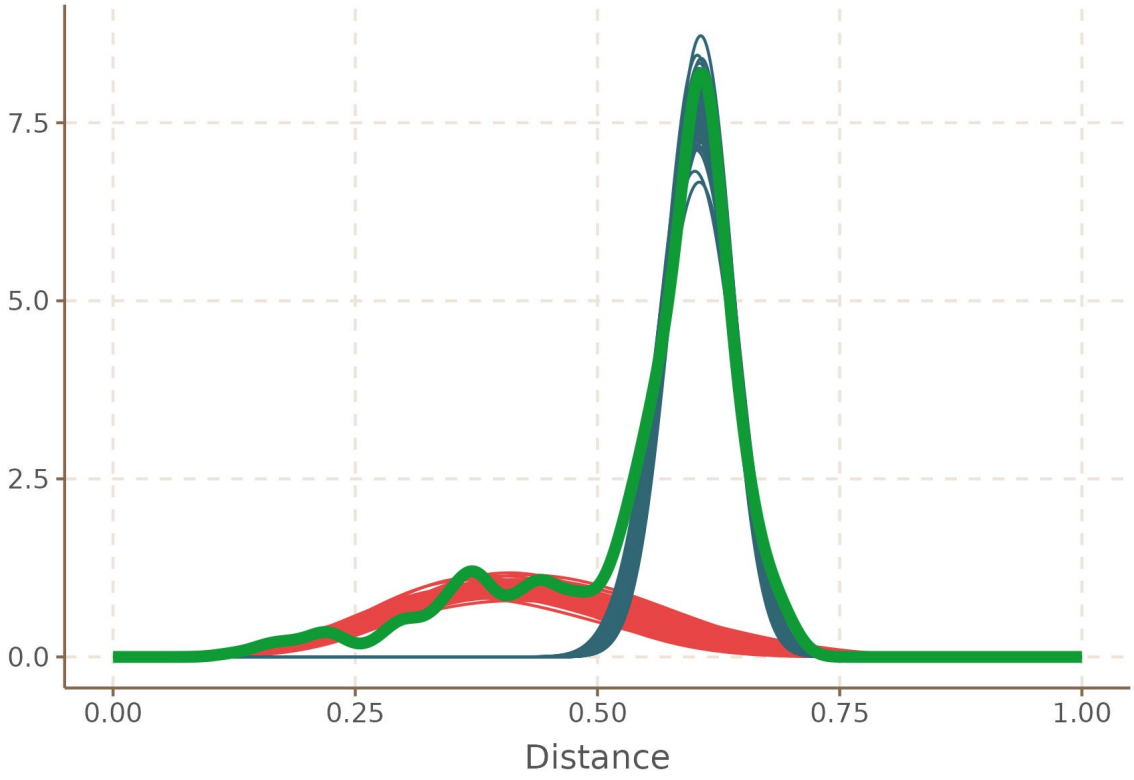

Figure 6. Observed distribution of grid-sample (distances using substitution costs) in green, and the two theoretical distributions estimated by the model M2 in red and blue.

From this result it is easy to see that while there is no clear cut-off point for the distance between two variants giving us certainty that we are dealing with either a pair of dialects or a pair of languages, it is still possible to observe an underlying categorical distinction between dialects and languages.

3.2 *Robustness checks*

One potential issue with our approach is the relatively small number of cognates in question. It is theoretically possible that the distributions we observe arise due to noise, and that it would even out if we were to include more words. To test this, we compare the distance matrix obtained using different numbers of words for smaller datasets for which we have more cognates. We focus on two subsets of our dataset: Romance doculects and Germanic doculects. For the Romance doculect dataset we have a total of 128 words and 38 doculects, and for Germanic 106 words and 58 doculects.

The setup of the experiment is as follows. We randomly sample N words from one of the datasets 100 times and then calculate the distance matrix using weighted Levenshtein distances. Subsequently we calculate the Steel-Penny distance from the resulting distance to the Glottolog tree for the relevant doculects. We repeat this process from *N*=1 to *N*=50.

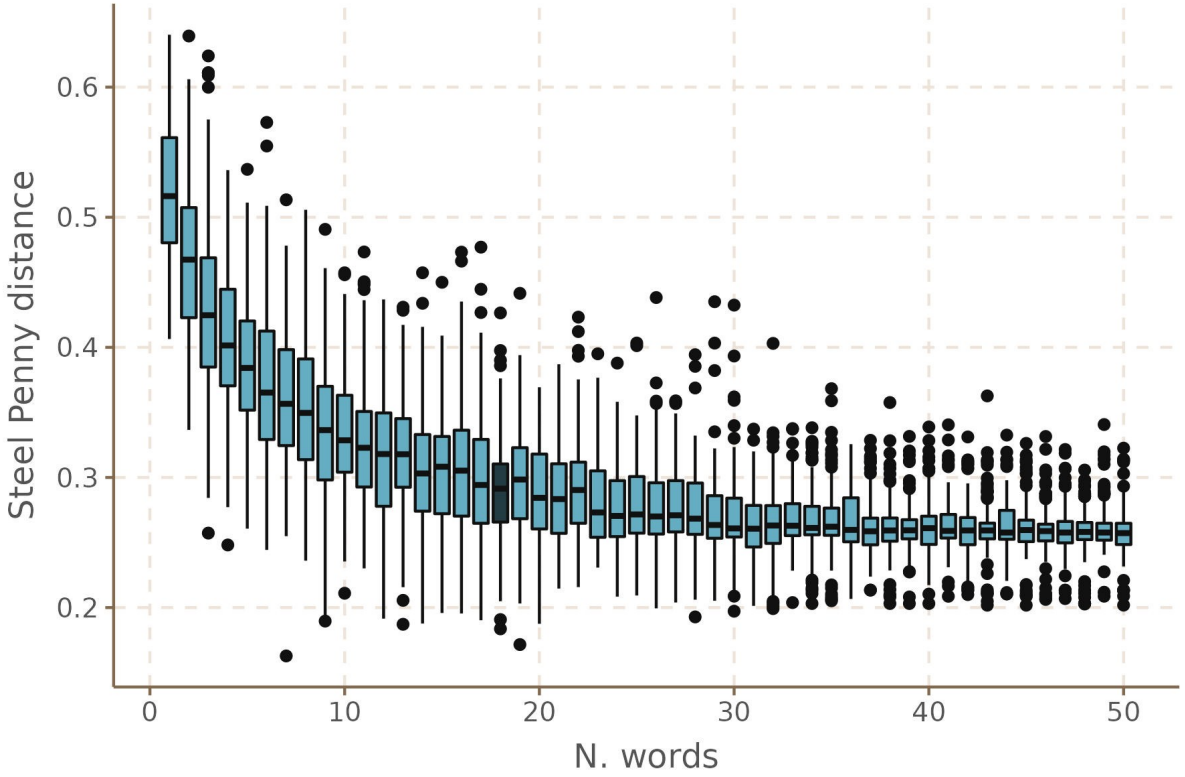

Figure 7. Distribution of Steel-Penny distances for different numbers of randomly sampled words for Romance doculects. *N*=18 is shaded in dark blue.

Figure 7 shows the results for the Romance doculects and Figure 8 the results for the Germanic doculects. We observe that the distance has relatively high variance for all *N*, and that it can vary considerably depending on the sampled words. The mean distance, however, stabilizes at around 30 sampled words,[13] and the mean distance for *N*=18 is very close to that mean. For the Germanic doculects we observe a very similar picture. The mean Steel-Penny distance stabilizes at more or less 18 words, and does not really improve after that. One

13 It is worth noting that in the original paper for ASJP the authors reach a similar conclusion.

feature of the Germanic lects is the large variability in distances depending on the random sample. This is likely due to contact effects throwing off the distance calculation.

Overall we observe that while having only 18 words in the whole dataset might be a source of some noise, and while we would expect results to improve with more words, it does not appear that the bimodal distribution is an artifact of having too few words.

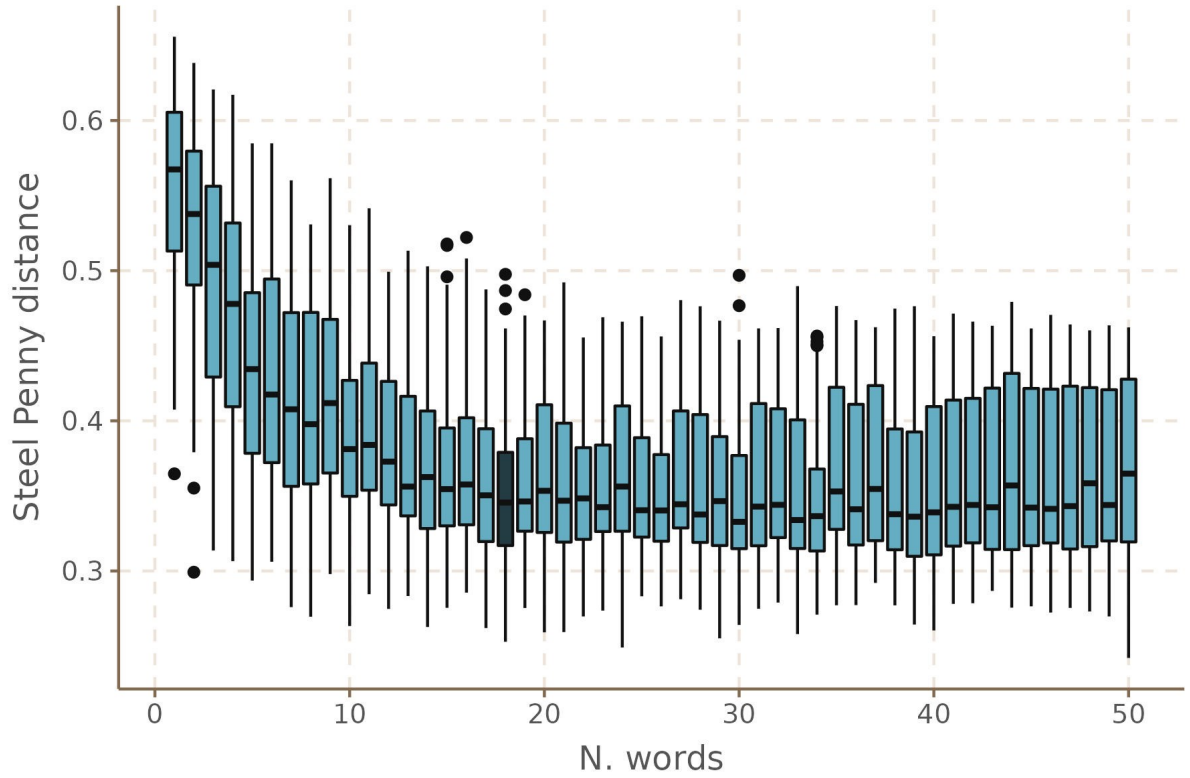

Figure 8. Distribution of Steel Penny distance for different numbers of randomly sampled words for Germanic doculects. *N*=18 is shaded in dark blue.

4. Discussion

In the preceding sections we hope to have established that pairs of language varieties tend to have relatively small or relatively large differences, with a shortage of in-between cases. We have suggested that the traditional label ‘dialects’ might be appropriate for the closer varieties and ‘language’ for the more distant ones. Although nothing dictates that we have to leave the comfort zone of abstract, statistical reasoning, it is tempting, of course, to ask how the findings resonate with practical aspects of language classification.

In this discussion we are going to address the following questions, where the first refers to the empirical cut-off point of 0.537 (0.527-0.547), which is the mean point at which the two component distributions in our mixture model intersect (see section 3.1 and Figure 8):

- Viewed as a non-negotiable 'iron curtain', how would an application of this cut-off compare with the separation of speech variants into dialects and languages implied by the ISO 639-3 standard?
- Is there some other classification procedure using our lexical-phonological distances which might produce a result closer to the ISO 639-3 standard?

We now go on to address these questions, but urge the reader to keep in mind that we are not viewing the ISO 639-3 as a gold standard against which to judge the quality of our results that were based on analyses of distance distributions[14]—rather, we are simply curious to see how they relate to one another.

The simplest way to compare the ISO 639-3 code classification to the classification obtained using the 0.527-0.547 Levenshtein distance cut-off is to count the intersection of lects that belong to same or different languages according to ISO 639-3 and lects that have distances greater than 0.547 or smaller than 0.527 or are in the 'gray zone' of 0.527-0.547. These counts are presented in Table 7.

Table 7. Numerical comparison of the number of pairs of lects in the six categories that emerge when crossing the ISO 639-3 classification with a classification according to a distance cut-off of 0.537 ± 0.01.

ISO 639-3 \ distance (D)	$D < 0.527$	$0.527 \leq D \leq 0.547$	$D > 0.547$
Same	732	2	1
Different	7074	1242	20,352

14 For the purposes of these explorations we only use the distances that take phoneme substitution costs into account.

The distribution of numbers in Table 7 clearly shows that the distance cut-off is comparatively speaking a lumper and the ISO 639-3 standard comparatively speaking a splitter when it comes to grouping together lects in what we may conveniently, for the purpose of the present discussion, call ‘languages’. There is only one pair which, according to the distance classification, consists of two different languages whereas it consists of different dialects according to the ISO classification. This pair is Occitan [oci]: Auverngat dialect of the Neschers region - Gascon dialect of the Val d’Aran. And only two pairs are in the gray zone with regard to distance but ‘same’ with regards to ISO-code. On the other hand, 90.6% of the pairs that would be considered dialects by the distance criterion (D < 0.527) are different languages according to the ISO-criterion. In order to explore the limits of what would be considered same languages according to the distance criterion, Table 8 offers the raw data for three pairs of lects that are just below the 0.527 limit.

Table 8. Most distant ‘dialects’ according to the distance criterion

ISO 396-3	lav	pol	gle	eng	mwl	src
name language	Latvian	Polish	Irish	English	Catalan	Sardinian
name lect	Latvian: Std	Polish: Mazovian	Cork: Ballymakery	Rossendale	Catalan	Sardinian N.: Monti
one	vĭjɛnts	jɻdan	eːn	wɒn	ʔuˑn	u̯ːnʊ
two	dĭˑvĭ	dva	doː	tʰʊuˑ	dọˑs	duˑɔ̯ˑzɔ̯
three	trɨːs	tʂɪ	tʰɾʲiː	θrɪiˑ	t̪ʰrə̯ːs	tə̆rɻːzɛ
four	tʃɛtrɪ	ʂtə̯rɪ	kʲʰæhʊɻ	fọˑəˑ	kʰwɐtrə	bɑ̯ttərɔ̯
five	pĭ̯jɻtsɪ	pçɐ̯ɲtɕ	kʰuˑɻ̍kʲ	fäˑɪv	siˑŋkx	kʰiˑmbɛ
six	sɛʃɪ	sɛ̨ɕtɕ	ʃeː	sɪks	siˑs	sɻːzɛ

seven	sɛptʂɲɪ	ɕʂdʂm	ʃaxt	sɛvə̆n	sɛ̟ːt̪ʰ	sɛ̟ˑttɛ
eight	hɑ̟stuɻnɪ	wɔɕʂm	ʊ̟xt	ɛˑɪt	voˑiţ̪ʰ	ɔˑttɔ
nine	dɛvʂɲɪ	dʑɛvɪɲtɕ	nʸeː	näˑɪn	nʌːu̟	nɔːɛ
ten	dɛsmɑ̟t	dʑɛ̟ɕĩɲtɕ	dɛ̟	tʰɛn	dɛ̟ːu̟	dɛ̟ːɣɛ
hundred	sɪmts	stɔ	kʲʰiɐd̪	ʊndɹəd	sɛ̟ːnţ̪ʰ	kʰɪːntu
tongue	mɛːɬɛ̟	jɛ̨ɯ̃zɪk	tʰaŋə	tʰɒŋg	ʎɛ̟ˑŋgo	liˑmba
name	vɑ̟ːɾts	imɲɛ̃	anʲɪmʲ	nɛ̟ːm	nɔ̟ːm	nɔ̃ːmɛ̃nɛ̃
wind	vɛːĭʃ	ʑɛ̟tʂ	fʲad̪	wɪnd	vɛ̟ːnţ̪ʰ	ɪːntʰʊ
salt	sɑ̟ːɬs	sʊ̟l	salʸən	sɒɬt	sa̟ːɬ	sa̟ːlɛ
grain	zʂrnʂs	ʑærka	gɹɑːn	kʰɔːɹn	gɾaː	ɻaːnʊ
young	ja̟ŏnts	mwɔdɪ	oːg	jʊŋg	dʒɟːvə̟	dʒɔːvɐnʊ̃
full	piːɬnɛ̟	pɛ̟wna	lʸɑːn	fʊɬ	pɬɜ̟ː	pʰjeːnʊ

While perhaps initially surprising, it is understandable that pairs such as the ones in Table 8 will occur given the limited set of words available and some potential reasons for similarity out of the ordinary relating to contact and shared retentions. The particular Polish variant is the one among several variants of Polish which is geographically closest to Latvia; Irish is in contact with English; Catalan and Sardinian may be united through Romance retentions.

Other illustrative cases are offered in Table 9. The first pair (Irish lects) is the one which has the smallest distance of all pairs in the data. In contrast to the cases in Table 8, we here see a pair of lects where the members are clearly extremely similar, with only a handful of phonetic differences in a couple of lexical items. The second pair represents a middle-of-the road case. It has a distance of 0.41, which lies very close to the mean of the dialect distribution. According to the ISO-codes, it would be classified as two different languages,

but using our method they are grouped together. Finally, the third pair is the pair having the greatest distance (0.556) among all the pairs which, according to ISO 396-3, should be classified as dialects.

Table 9. Additional examples of dialect-language distinctions. (1) the two closest varieties in the dataset, (2) two dialects close to the mean of the dialect distribution, (3) two lects which, according to ISO 396-3 should be dialects, but have the greatest distance among pairs representing one and the same ISO-code.

ISO 396-3	gle	gle	wep	gsw	oci	oci
name language	Irish	Irish	Wesphalien	Allemanic	Ocitanian	Ocitanian
name lect	Gweedore-Carrick	Tory Island	Westphalia: Münsterland	Liecht.: Oberland	Gascon: Val d'Aran	Auvergne: Neschers
one	eːn	eːn	e̞ɪnə	ʔäːs	yˑũŋ	vʌ̃ˑ
two	dɑː	dɑː	t̪βe̞ːə	tsvoː	dyːs̪	du
three	tʰɾ̪iː	tʰɾ̪iː	dɾeɪə	dryː	t̪ɾeːs̪	tʰχi
four	kʲʰɛhəɻ̪	kʲʰæhæɻ̪	fe̞ːɾə	fiːɛ	kwa̤ˑte	kʰɐtχɤ̆
five	kʰuɪkʲ	kʰuˑɪc	fiːʊə	fyːf	s̪iˑŋ	ʃĭˑ
six	ʃeː	ʃeː	sɛsə	sɪk̟s	s̪iˑe̞s̪	ʃei
seven	ʃaxt	ʃaxt	siˑəm	siˑb̥ɐ	s̪ɛːtˀ	sɪ
eight	ɑxt	ɑxt	ʔaxt	ʔäˑχt	uweit̪ʰ	ɣ̊əˑ
nine	nʸiː	nʸiː	niəgn̥	nyː	na̤ˑu	nəˑ

ten	dʒæjʰ	dʒæj	t̪aɪn	tsehɐ	d̪e̞t̪ʂ	ɟɪˑ
hundred	kʲʰɛːd̪	kʲʰɛːd̪	hʊnɔt̪ʰ	hundətʰː	ʂeˑn	sɛ̃ˑ
tongue	tʃaŋi	tʃaŋi	t̪ʰʊŋən	tsu̪ŋˑa̪	leˑŋgwʌ	iʌiŋgɻ̟
name	aɲɪmʲ	aɲɪmʲ	næːm	namä	no̞ˑm	nuˑ
wind	fʲaˑd̪	fʲad̪	viːntʰ	vi̞ntʰ	beˑnt	vɛ̃ˑ
salt	salʸən	salʸən	sɔltʰ	säːlts	ʂa̪ˑuˑ	soˑ
grain	gɻaˑɲ	gɻan	kʰœɾn̩	kʰɒʁn	gɾaː	gχoˑ
young	ɑːg	ɑːg	jʊŋg̊	jʊŋ	dʒŭwe̞n	dʰʊinə̆
full	lʸɛːn	lʸæːn	fʊl	fol	pleˑŋ	plʸə

Regardless of how to explain cases such as those in Table 8, it is clear that a simple ‘one size fits all’ approach according to which varieties below some cut-off are dialects of one another and varieties above some cut-off are different languages is not tenable. For instance, the ‘Greater Poland’ variant of Polish has a distance to Standard Latvian of 0.5426, exceeding the cut-off, while its distance to Mazovian Polish is 0.2268. Thus, there is a conflict between on the one hand the ‘same language’ status of Mazovian Polish-Standard Latvian and Mazovian Polish-Greater Poland Polish and on the other hand the ‘different language’ status of Standard Latvian-Greater Poland Polish. The way to avoid such conflicts is to define clusters rather than pairs when applying a language vs. dialect criterion. We now turn to this.

There are different approaches to clusterization. For any approach we will want to compare the clusters obtained with clusters as defined by the ISO 639-3 standard. We do this using the Normalized Information Distance (NID),[15] which is an entropy-based measure equal to 1 – the normalized mutual information between a set of clusters. It produces values in the 0 to 1 range.[16]

15 Some alternatives are the split/join metric, Rand Index or Variation of Information index (see Meilă 2007 for a discussion of different metrics).
16 The NID() function of the aricode R package (Chiquet et al. 2020) shows how this metric is defined .

One approach to building clusters from the distance matrix is to use a hierarchical clustering method and then cut the tree at some height.[17] We can choose the 0.537 cut-off point which emerged from our data as the height at which to cut the tree. We expect that two lects with a distance below this point will likely belong to the first component (the 'dialect component') in the mixture of two distributions, and that lects with distances above this point will likely belong to the second component (the 'language component'). This approach leads to 13 different clusters. It is already clear from a comparison of the corresponding number of clusters in the ISO 639-3 classification, which is 72, that the latter is more inclusive and rather different. The relatively high NID value of 0.427 is a manifestation of the differences. An example of one of the concrete differences that emerge through inspection is the treatment of several Romance varieties that are merged in the distance-based clusterization but distinguished in the ISO 639-3-defined clusters.

We could alternatively aim at a better approximation to the ISO 639-3 classification, minimizing the NID, while still using tree-cutting with a certain cut-off. Using an optimization technique to achieve this goal results in an optimal cut-off point at 0.327 and a NID of 0.121. Using this cut-off point already gets us very close to the ISO 639-3 classification.

These tree-based approaches assume that all clusters must be cut at the same height, and therefore require us to pre-specify a height at which to cut the tree. An alternative to building clusters which does not rely on cut-off points (or at least not to the same extent), is the following, which we might call a nearest neighbor approach. For each lect L_0, we find its nearest neighbor L_1, i.e. the lect with the smallest distance to it, and add L_1 to the cluster of L_0, C_0. We repeat this process for L_1 and find its nearest neighbor L_2 and add this to C_0. The next nearest neighbors are added until we reach a loop, i.e. until we find an L_n already in C_0. This process is repeated for all lects, and then all clusters sharing one or more lects are joined together. This process results in 55 clusters when continued until it contains all lects. While this method has the advantage of being agnostic to cut-off points, it has the drawback that it necessarily groups 'isolated' lects together with some other lect(s). If, for example, a single variant of Basque was included in our dataset, it would necessarily end up together with one or more other lects in some cluster. To avoid this, one could again resort to a distance cut-off, stipulating a limit on the distance between nearest neighbors, and here the 0.537 distance

17 These operations can be carried out using the base R functions hclust() (default settings) and cutree().

suggests itself as an independently motivated option for such an upper limit. Since we did not have ‘isolates’ in our dataset, however, we did not have to introduce this amendment. In practice the NID distance between a clusterization obtained through the ‘nearest neighbor’ approach and the ISO 639-3 classification is 0.163, which is remarkably close considering that ISO 639-3 did not play any role in the classification process.

In Table 9 we offer pairwise comparisons of the different clusterizations with each other and with ISO 639-3, here ‘C-ISO’ for short. The clusterizations obtained by cutting the tree at 0.327 and 0.537 are called C-0.327 and C-0.537, respectively, and the ‘nearest neighbor’ approach is called C-N.

Table 9. NID distances between different clusterizations

Clusterizations compared	NID
C-0.537 vs. C-ISO	0.427
C-0.537 vs. C-0.327	0.405
C-0.537 vs. C-N	0.408
C-ISO vs. C-N	0.163
C-0.327 vs. C-ISO	0.121
C-0.327 vs. C-N	0.126

The explorations of this section have shown that using a cut-off found by studying distance distributions does not necessarily lead to a classification that has a good fit with the ISO 639-3 classification of lects, at least not in a straightforward way; but it could be relevant as an upper bound on distances among members of cluster obtained in some such way as the ‘nearest neighbor’ approach tried out here. Given the data limitations we do not venture into more extensive explorations of possible practical implications of our findings for a dialect-language classification. Not only is the set of doculects available limited, so is the set of words available for each lect. It is doubtful that more than very general observations can be

based on distances computed from just 18 lexical items. But a future study drawing upon the ASJP database and using a combination of clustering and a distance cut-off may lead to a meaningful and useful classification of speech varieties throughout the world into languages and dialects.

Finally, there are a couple of limitations of the present study which are worth mentioning. Although we show that the relatively small number of words is likely not affecting our results in major ways, a study based on a larger sample would be even more decisive. Another issue we were not able to address in this study, is the question of whether our results would generalize to other language families and socio-geographic situations. This is, of course, a question which needs to be tested empirically, but we are not aware of datasets of comparable quality for other parts of the world. The Sound Comparisons dataset does provide data for some Andean varieties, but this is considerably less detailed and covers a smaller area. Alternatively, as we show, the ASJP transcription leads to distance measures that are not very different in quality from those based on the far more detailed phonetic transcription system used by the Sound Comparisons data, so similar grid-based studies could be carried out for other parts of the world using the ASJP database, although metadata relating to language locations would have to revised.

5. Conclusion

In an essay called ‘Some issues on which linguists can agree’, a prominent linguist claimed, on the one hand, that “[t]here is no clear or qualitative difference between so-called ‘language-boundaries’ and ‘dialect-boundaries’” while, on the other hand claiming that the number of languages in the world can be estimated at 4000-5000 even if “no precise figure is possible” because of the uncertainty referred to in the other claim (Hudson 1981: 336). While some linguists perhaps agree that one can roughly estimate the number of languages in the world while also agreeing that this number is essentially unknowable, others try to be more methodological. Hammarström (2016) refers to lack of mutual intelligibility as a criterion for distinguishing languages and ventures an estimate of 6,500 languages (including known extinct languages).

This paper calls into question the claim that there is no identifiable transition between dialects and languages. More specifically, we have shown that the general results in Wichmann (2019), which support the existence of an identifiable transition, can be replicated

using a different, and in some regards perhaps more appropriate, dataset, consisting of European language varieties. We have shown that the distribution of distances in our dataset also clearly does follow a bimodal distribution and that, similarly to the findings of Wichmann, the two underlying distributions intersect at a distance (or similarity) close to fifty percent.

At present we can think of two explanations for the bimodal distribution of distances. The first was voiced by Wichmann (2019: 830), who suggests that the phenomenon of mutual intelligibility should be taken into account. He speculates that the proposed language-dialect distance cut-off may correlate with a point where mutual intelligibility between two lects A and B is so low that speakers of A and B will tend no longer to use their individual lects when communicating with one another. The lects would then be less likely to influence one another, speeding up their differentiation, much as when a spring is released. The implied argument here is that a researcher would be relatively unlikely to record two lects in the precarious, transitional state. Clearly, more research on mutual intelligibility is needed to investigate this suggested. Another explanation for the bimodal distribution is the one alluded to in the title of this paper. We have introduced ‘the half-way similarity avoidance rule’ as a term for describing the phenomenon that language varieties apparently tend not to be around half-way similar with respect to the kind of lexical-phonological data looked at here and in Wichmann (2019). If there really is some agency-driven avoidance going on, which is of course rather speculative, the underlying cause must be of a sociological nature. People belong to social groups, and groups that one caters to (in-groups) are often partly defined through non-membership in other groups (out-groups) (Tajfel 1974). Thus, it is natural to maintain a relation to one's peers which is distinctly closer than the relation to members of other groups, rather than a relation which is somewhere in between two groups. We hypothesize that the ‘rule’ may reappear using other areas of language than phonology and the lexicon, and perhaps even in other cultural manifestations. At the same time we are well aware that these ideas are quite immature and that they bring us into areas of research (sociology, social psychology, anthropology) that we need to get better acquainted with before attempting to develop them any further.

The robust, non-speculative result of this paper is that language variant pairs tend to fall in two clusters, even though there is a gray area in between, which in practice sometimes makes it difficult to distinguish them. Although this choice is not strictly necessary, we

choose the labels ‘dialect’ and ‘language’ for the clusters, in order to relate our findings to these traditional categories and to help define the latter better. But the empirical definitions of dialects vs. languages depend on the data used, and for the practical purpose of cataloging languages vs. dialects, even for the geographical region considered in the present paper, the data at our disposal were likely insufficient.

There is plenty of room for more validation of an identifiable transition between dialects and languages, for further explorations of whether such a transition may ultimately be grounded in mutual intelligibility, and for also exploring sociological explanations. There are also prospects of practical applications, including the development of a new take on the distinction between languages and dialects at a world-wide scale.

Supplementary materials

The supplementary materials contain all the code used for the experiments, distances, models and plots. It also contains the distance matrices, substitution cost matrices, trees, and other materials necessary for replicating the present study. The supplementary materials can be found at https://doi.org/10.5281/zenodo.7752997 .

References

Bakstrom, Peter C. and Carla F. Radloff. 1992. *Sociolinguistic survey of Northern Pakistan 2. Languages of northern areas*. Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics.

Beniamine, Sacha. 2017. Un algorithme universel pour l'abstraction automatique d’alternances morphophonologiques. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)* (Vol. 2, No. 2017).

Beniamine, Sacha and Matías Guzmán Naranjo. 2021. Multiple alignments of inflectional paradigms. In *Proceedings of the Society for Computation in Linguistics 2021*. 216-227.

Brown, Cecil H., Eric W. Holman, and Søren Wichmann. 2013. Sound correspondences in the world's languages. *Language* 89.1:4-29.

Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1):1–28. https://doi.org/10.18637/jss.v080.i01

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1):1–32. https://doi.org/10.18637/jss.v076.i01

Chiquet, Julien, Guillem Rigaill, and Martina Sundqvist. 2020. aricode: Efficient computations of standard clustering comparison measures. R package version 1.0.0. Url: https://CRAN.R-project.org/package=aricode

Dellert, Johannes, and Gerhard Jäger. 2017. NorthEuraLex (Version 0.9). Tübingen: Eberhard-Karls University Tübingen.

Downey, Sean S., Hallmark, Brian, Cox, Murray P., Norquest, Peter, & Lansing, J. Stephen (2008). Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. *Journal of Quantitative Linguistics*, *15*(4), 340-369.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(2):389–402. https://doi.org/10.1111/rssa.12378.

Greenhill, Simon. 2018. treemaker: A Python tool for constructing a Newick formatted tree from a set of classifications. *Journal of Open Source Software* 3(31), 1040. https://doi.org/10.21105/joss.01040.

Hammarström, Harald. 2016. Linguistic diversity and language evolution. *Journal of Language Evolution* 1:19–29. doi: 10.1093/jole/lzw002

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. Glottolog 4.5. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5772642 (Available online at http://glottolog.org, Accessed on 2022-01-05.)

Heggarty, Paul, Aviva Shimelman, Giovanni Abete, Cormac Anderson, Scott Sadowsky, Ludger Paschen, Warren Maguire, Lechoslaw Jocz, María José Aninao, Laura

Wägerle, Darja Dërmaku-Appelganz, Ariel Pheula do Couto e Silva, Lewis C. Lawyer, Ana Suelly Arruda Câmara Cabral, Mary Walworth, Jan Michalsky, Ezequiel Koile, Jakob Runge & Hans-Jörg Bibiko. (2019). Sound comparisons: A new resource for exploring phonetic diversity across language families of the world. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Canberra, Australia 2019*. Australasian Speech Science and Technology Association Inc. 280-284. [The online resource, at https://soundcomparisons.com/#home, was accessed 2021-07-13].

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52:841–875.

Holman, Eric W., Christian Schulze, Dietrich Stauffer, and Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguistic Typology* 11(2):395–423.

Hudson, Richard. 1981. Some issues on which linguists can agree. *Journal of Linguistics* 17:333–343.

Kilani, Marwan. 2020. FAAL: A Feature-based Aligning ALgorithm. *Language Dynamics and Change* 11:30–76. https://doi.org/10.1163/22105832-01001300.

Kumar, Sudhir, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547–1549.

Legendre, Pierre and François-Joseph Lapointe. 2004. Assessing congruence among distance matrices: single-malt Scotch whiskies revisited. *Australian & New Zealand Journal of Statistics* 46:615–629.

List, Johann-Mattis. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, 9-12. Valencia: Association for Computational Linguistics.

List, Johann-Mattis, and Robert Forkel. 2022. LingPy. A Python library for quantitative tasks in historical linguistics [Software library, version 2.6.9]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://pypi.org/project/lingpy.

Meilă, Marina. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98 (5): 873–95. https://doi.org/10.1016/j.jmva.2006.11.013.

Mortensen, David R., Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, Lori Levin. 2016. PanPhon: A Resource for mapping IPA segments to articulatory feature vectors. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan, December 11-17, 2016.

Oelrich, Oscar, Shutong Ding., Mans Magnusson, Aki Vehtari and Mattias Villani (2020). When are Bayesian model probabilities overconfident?. *arXiv preprint arXiv:2003.04026*.

Paradis, Emmanuel and Klaus Schliep. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:525–528.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53(1-2):131–147. doi: 10.1016/0025-5564(81)90043-2.

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4):406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Sand, Andreas, Morten K. Holt, Jens Johansen, Gerth Stølting Brodal, Thomas Mailund, Christian N. S. Pedersen. 2014. tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics* 30:2079–2080. doi: 10.1093/bioinformatics/btu157.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane* 35:335–357.

Smith, M. R. 2019. Quartet: comparison of phylogenetic trees using quartet and split measures. R package version 1.2.2. doi:10.5281/zenodo.2536318.

Smith, M. R. 2020. TreeDist: Distances between phylogenetic trees. R package version 2.3.0. Comprehensive R Archive Network. doi:10.5281/zenodo.3528124.

Steel, Mike A. and David Penny. 1993. Distributions of tree comparison metrics—some new results. *Systematic Biology* 42(2):126–141. doi: 10.1093/sysbio/42.2.126.

Tajfel, Henri. 1974. Social identity and intergroup behaviour. *Social Science Information/sur les sciences sociales* 13: 65–93.

Van Rooy, Raf. 2020. *Language or Dialect? The History of a Conceptual Pair*. Oxford: Oxford University Press.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5):1413–32. https://doi.org/10.1007/s11222-016-9696-4.

Voegelin, Carl F. and Zellig S. Harris. 1951. Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society* 95:322–329.

Wichmann, Søren. 2019. How to distinguish languages and dialects. *Computational Linguistics* 45:823–831.

Wichmann, Søren and Harald Hammarström. 2020. Methods for calculating walking distances. *Physica A* 540, 122890. https://doi.org/10.1016/j.physa.2019.122890

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown, editors. 2018. The ASJP Database (version 18). http://asjp.clld.org/.