BRILL

LANGUAGE DYNAMICS AND CHANGE 9 (2019) 104–126

brill.com/ldc

Areal dependency of consonant inventories

Dmitry Nikolaev
The Hebrew University of Jerusalem
dnikolaev@fastmail.com

Abstract

This paper discusses the impact of linguistic contact on the make-up of consonantal inventories of the languages of Eurasia. New measures for studying the importance of language contact for the development of phonological inventories are proposed, and two empirical studies are reported. First, using two different measures of dissimilarity of phonemic inventories (the Jaccard dissimilarity measure and the novel Closest-Relative Cumulative Jaccard Dissimilarity measure), it is demonstrated that language contact—operationalized as languages being connected by an edge in a neighbor network—makes a significant contribution to between-inventory differences when phylogenetic variables are controlled for. Second, a novel measure of the exposure of a language to a particular segment—the Neighbor-Pressure Metric (NPM)—is proposed as a means of quantifying language contact with respect to phonological inventories. It is shown that addition of NPM helps achieve higher prediction accuracy than using bare phylogenetic data and that distributions of different consonants display a different degree of dependence on language-contact processes. Finally, more complex models for predicting consonant inventories are briefly explored, demonstrating the presence of complex non-linear relationships between inventories of neighboring languages.

Keywords

consonant inventories – spatial dependencies – statistical modelling – neighbor graphs

1 Introduction

The search for the factors shaping the structures of phonological inventories of the world's languages has been one of the primary goals of phonological typol-

 | DOI:10.1163/22105832-00901001

ogy since the work of Trubetzkoy (1939) and Hockett (1955). Early structuralists' attempts at formulating the laws of inventory structures in terms of a minimal number of oppositions gave rise to a rich literature on phonological-feature theory aiming at enumerating a minimal set of compositional atomic properties of phonemic segments necessary and sufficient to describe the world's phonologies (Clements 2003; Mielke 2008) and at producing generalizations about structures of phonological systems as described using such feature sets (Dunbar & Dupoux 2016).

Non-formalist attempts at inventory-structure modelling were markedly less numerous and were nearly exclusively concentrated on vowels, presumably since those can be thought of as inhabiting a continuous low-dimensional space of formants amenable to mathematical modelling (Liljencrants & Lindblom 1972; Stevens 1989; Cotterell & Eisner 2017).

This work on the consonants and consonant-and-vowel inventories was mostly descriptive in nature. Starting with the seminal monograph by Maddieson (1984), researchers have put a great deal of effort into discovering statistical tendencies in the distributions of different types of segments and their correlations based on cross-linguistic samples, cf. a recent overview by Gordon (2016).

A series of publications appearing in the last several years suggests that there are extra-linguistic correlates possibly influencing the structure of segmental inventories of the world's languages, such as distance from Africa, arguably negatively correlated with phonological complexity (Atkinson 2011 cf. the critique in Jaeger et al. 2011); average annual temperature and sexual freedom (positively correlated with the prominence of sonority in an inventory Fought et al. 2004; Ember & Ember 2007); population size (positively correlated with phonological complexity: Hay & Bauer 2007; Trudgill 2011; Atkinson 2011; Wichmann et al. 2011; Nettle 2012, cf. objections in Donohue & Nichols 2011 and Moran et al. 2012); altitude at which the language is spoken (possibly conducive to the development of ejective sounds, see Everett 2013), and several others. A general review of correlation studies is presented in Ladd et al. (2015).

These results enrich our understanding of the properties of human language, but they do not go a long way towards answering the following basic question: to what extent are the world's languages shaped by non-universal factors? The importance of this question is hard to overestimate. Ian Maddieson writes: "As an illustration, the modal number of vowel qualities in language inventories is shown to differ for the set of African languages in UPSID from those of other continents." (Maddieson 1991: 193) This effectively implies that, for the purpose of linguistic sampling, all languages outside Africa form a unit, samples from which cannot be considered independent. Maddieson does not

provide a statistical analysis which could substantiate such a claim,[1] but the possibility itself is troubling for many kinds of linguistic sampling procedures.

Generally speaking, we are presented with a case of duality: if we want to predict particular properties of particular languages we need to have a good grasp of the contribution of the same factors we need to control for when aiming at general statements about the typology of languages.

Two factors are usually held accountable for non-randomness in the distribution of linguistic variables in the world's languages: linguistic phylogenetics, or inheritance, and contact phenomena.

Linguistic phylogenetics is a very complicated and hotly debated subject (Chang et al. 2015; Wichmann 2017), but in practical typological analyses it is usually considered non-controversial: most of the world's languages are assigned to one of the several relatively well-established large families (phyla), which in most cases have an agreed-upon set of large subgroups (genera). These data are available in repositories such as Ethnologue[2] and Glottolog[3] and are routinely included in quantitative typological datasets as control variables. Some work has been done to directly estimate the impact of phylum membership on the distribution of linguistic variables (Bickel 2013).

Contact phenomena are much less straightforward. As Ladd et al. put it: "Unfortunately, for purposes of conducting correlational analyses, quantifying contact is even more difficult than quantifying genealogical relatedness. (A further complication arises from the fact that, by the nature of language splits, languages in contact also tend to be related.)" (Ladd et al. 2015: 230–231) Moreover, it has long been known that non-neighboring languages may display non-trivial similarities as a consequence of belonging to the same linguistic area (Nichols 1992; Thomason 2000). A long list of linguistic areas have been proposed in the literature, together covering a good share of the Earth's land mass (e.g., mainland South East Asia (Enfield 2005), the Macro-Sudan belt (Güldemann 2008), India (Emeneau 1956), the Balkans (Friedman 2000), the Circum-Andean region (Michael et al. 2014), etc.—the notion of linguistic area, however, is far from unproblematic, cf. Campbell 2006, 2017). In most cases, trying to account for possible local disturbances, typologists have used very big landmasses roughly equivalent to continents as control variables (Dryer 1989). However, some proposed linguistic areas, such as the Pacific Rim (Bickel & Nichols 2006) or Beringia (Fortescue 1998), cut across even these macro regions.

1 It may be remarked that to prove such a bottleneck scenario there should be several comparable areas in Africa, and not just one.

2 https://www.ethnologue.com/.

3 http://glottolog.org/.

Theoretically, it is possible to account for complex areal structures by assigning languages to a succession of areas of predefined magnitude. Thus, a language from the western coast of South America can simultaneously participate in (i) the Circum-Andean linguistic area, (ii) the South American linguistic area, and (iii) the Pacific-Rim area. However, several intersecting linguistic areas can potentially exist on any level of geographical magnitude, making any analysis based on a fixed number of area variables problematic.

It seems that there are two ways to overcome the problem of areality. The first way is to dispense with the idea of independently pre-defined linguistic areas, statistically recover areality patterns in the distribution of different linguistic features, and then use the results of this procedure for sampling and hypothesis testing (Daumé III 2009).

The second way is to make a simplifying assumption that it is possible to devise a metric measuring the pressure exerted on a given language by its neighbor languages nudging it towards acquiring or preserving a feature of interest. If, leaving aside the hypothesis that different linguistic features have different diffusion rates in different regions (Wichmann & Holman 2009), we can disregard wider areal structures, we will be left with much simpler statistical models involving fewer nominal predictors, which, if numerous or multileveled, demand prohibitive amounts of data to achieve statistically significant results.[4] This is the approach that will be further explored in this paper.

It must be noted that in order to substitute a numeric neighbor-pressure variable for a nominal macro-areal one we must have genuine local information, at least partly directly reflecting the contact history of the languages in the sample. Geographically sparse samples, such as those used in WALS chapters (Dryer & Haspelmath 2013), are not suited for this task as they are not spatially dense or even spatially uniform. The only type of linguistic data for which spatially dense samples for large regions is available at the moment is phonological segmental inventories. Most results in phonological typology were achieved using Ian Maddieson's balanced sample (Maddieson 1984; Maddieson & Precoda 1992); however, in recent years a series of phonological databases including PHOIBLE (Moran et al. 2014), SAPhon (Michael et al. 2015), and the Database of Eurasian Phonological Inventories (EURPhon) (Nikolaev et al. 2015) have provided dense samples of phonological data for several macro regions, especially Eurasia and South America.

4 Importantly, even if we do not make such an assumption, it will be possible to account for differences between regions by assigning some independently computed 'local-intensity-of-contact' values to each region and then using these values to weight the language-contact-intensity variable.

The aim of this paper is to use these data in order to assess the dependence of consonant inventories of Eurasian languages on the inventories of neighboring languages.

In § 2, I introduce the language data and the neighbor network of Eurasian languages used to operationalize the notion of language contact in the reported studies.

In § 3, using pairwise whole-inventory comparisons, I show that language contact played a significant rôle in shaping consonant inventories of Eurasian languages, clearly detectable over and above that of phylogenetic inheritance and even possibly obliterating the effect of phylogeny on the level of the phylum.

In § 4, a new local density measure—the neighbor-pressure metric (NPM)—is proposed. This measure is then used to model distributions of consonants in languages of Eurasia. It is shown that distributions of different segments show a different degree of dependence on language contact, viz. that some segments easily cross inter-genus and inter-phylum boundaries while others seem to be transmitted 'vertically'.

A series of models is then fit to predict distributions of consonants in the languages of Eurasia. It is shown that the addition of NPM to the model noticeably helps improve prediction accuracy, but that significantly better results can be achieved by allowing for complex non-linear dependencies between inventories of neighboring languages and by considering internal inventory structures.

Section 5 concludes.

2 Data and the neighbor graph

2.1 *Data*

Languages of Eurasia were chosen as a dataset for this study because we have access to a sample of languages from this region with the necessary degree of spatial density and sufficiently varied in terms of linguistic phylum and genus membership. For the purposes of this paper, a pooled dataset was constructed consisting of Eurasian data from PHOIBLE (regions 'Asia' and 'Europe' excluding languages of Indonesia, Malaysia, and Brunei) and all data from EURPhon, together comprising 481 languages.

All segments were parsed into an IPA-chart-based feature representation, i.e. /p/ is represented as {voiceless, bilabial, plosive} and /ʈʷ/, as {voiceless, retroflex, plosive, labialized}.

2.2 *The neighbor graph*

For analyses of dependence of linguistic variables on contact processes it has become a common practice to construct neighbor graphs of languages. In such a graph, languages are usually represented by points on a map and pairs of languages are connected by edges if the distance between them does not exceed a certain threshold, which is usually taken to be 1000 kilometers (Towner et al. 2012; Yamauchi & Murawaki 2016). After constructing such a graph, it becomes straightforward to consider only immediate neighbors of a given language or to construct another graph, where languages are connected if they belong to the same phylogenetic grouping and then compare distributions of features on these two graphs.

This approach assumes that languages are on average in contact with 'nearby' languages. However, large maximal distances between neighbors force researchers to postulate spurious cases of distal contact phenomena: 1000-edge graphs produce cases when language A 'influences' language B, even though it is separated from it by the territory of language C. Such cases are not, to my knowledge, discussed in the contact literature (except for cases of super-regional languages such as Arabic or Russian, which cannot be easily captured by any purely geography-based graph model), and it seems that it is safe to assume that languages engage in most intense contact with their immediate neighbors.

The latter notion is not strict, but it is easily formalized in the form of Delaunay triangulation of an array of points (Delaunay 1934). In this triangulation, three points are connected by edges if they form a triangle (or, in higher dimensions, a simplex) whose circumference does not include any other points.

An important property of Delaunay triangulation is that it always traces the convex hull of the points. This makes it easy to construct neighbor graphs for geographical points without actually computing a Delaunay triangulation itself: after converting latitudes and longitudes to 3D Cartesian coordinates (assuming that the Earth is spherical with its center at the origin), we can find the convex hull of the points and then delete the edges connecting languages separated by a geodesic distance that exceeds some predefined limit (for this study taken to be equal to 1000 kilometers following da Silva & Tehrani 2016 and Murawaki & Yamauchi 2018). Using this method, if languages A and B are separated by the territory occupied by language C they are extremely unlikely to be connected by an edge. Conversely, the upper limit makes sure that no spurious long-range contacts between languages on continental boundaries are assumed. The neighbor graph used in the analyses in this paper is shown in Figure 1.

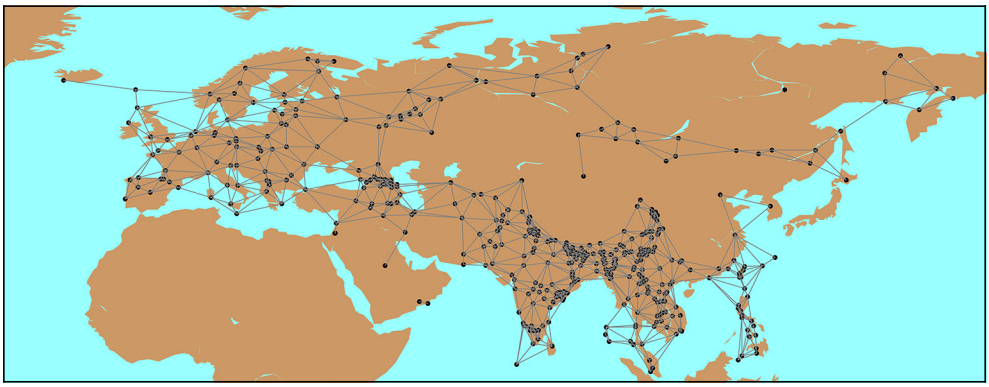

FIGURE 1 Neighbor graph of the languages of Eurasia analyzed in this study

This graph provides a formal definition of neighboring languages—as those separated by one edge—and allows for a new interpretation of distance between languages as measured in 'degrees of separation'. Further discussion of the latter notion is deferred to § 4, where NPM is introduced.

3 Pairwise distances between inventories

Using the neighbor graph described in the preceding section and the phylogenetic metadata provided by PHOIBLE and EURPhon, it is possible to divide all the language pairs in the sample into six groups according to whether they are neighbors and belong to the same phylum and the same genus. The groups are described in Table 1.

Using a dissimilarity metric, we can then compute phonological differences in the language pairs and test for differences between different groups (with the null hypothesis that median distances in pairs from different groups will be the same). The key question is whether we can induce an ordering on these six groups and, if yes, how this ordering is determined.

Two dissimilarity metrics were used in this study.

The first one is the classical Jaccard dissimilarity:

$$\text{Jaccard}(\text{A}, \text{B}) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

where A and B are sets of phonemes found in the respective languages.

The Jaccard metric does not take into account a common diachronic process of phoneme splitting due to development of VOT distinctions or additional

TABLE 1 Classification of language pairs in the dataset

Coding	Neighbors	Same phylum	Same genus	# of language pairs
(–n; –p; –g)	–	–	–	97129
(+n; –p; –g)	+	–	–	1134
(–n; +p; –g)	–	+	–	13950
(–n; +p; +g)	–	+	+	1583
(+n; +p; –g)	+	+	–	890
(+n; +p; +g)	+	+	+	754

articulations, such as aspiration and glottalization. This leads to the fact that there is no reported difference between, on the one hand, a pair of inventories that share some core of segments with one of them having enriched this core by developing some additional articulation (a simplified example could look like /p, t, k/ vs. /p, t, k, pʰ, tʰ, kʰ/), and on the other hand, a pair of inventories sharing a common core beyond which the rest of the phonemes are not evidently related (/p, t, k/ vs. /p, t, k, ʔ, ʕ, χ/). A real-life example is that of Serbian, which, unlike most Slavic languages, lacks the hard–soft contrast, but whose overall inventory structure is rather similar to and more or less 'derivable' from e.g. that of Russian. Consequently, it may be assumed that the Jaccard metric often overestimates differences between inventories.

In order to overcome this limitation, a new metric is proposed, Closest-Relative Cumulative Jaccard Dissimilarity (CRCJ). It is computed in the following way: given inventories I_1 and I_2 where each phoneme is described as a set of IPA features, for each phoneme p in I_1 we find its closest relative in I_2—that is, the phoneme in I_2 having the smallest feature-wise Jaccard dissimilarity from p—then sum these minimal Jaccard dissimilarities and finally repeat the procedure starting with I_2.[5] Formally, the metric can be defined as follows:

$$\mathrm{CRCJ}(I_1, I_2) = \sum\nolimits_{p \in I_1} \mathrm{Jaccard}(p, q) + \sum\nolimits_{r \in I_2} \mathrm{Jaccard}(r, s)$$

where $q = \arg\min_{x \in I_2} (\mathrm{Jaccard}(p, x))$ and $s = \arg\min_{x \in I_1} (\mathrm{Jaccard}(r, x))$.

5 For example, given the inventories /p, t, k/ vs. /p, t, k, pʰ, tʰ, kʰ/, we first add 0 for each segment from the first inventory, as they all have direct counterparts in the second one, then add 0 for /p, t, k/ from the second inventory, and finally add 1–3/4 for each of /pʰ, tʰ, kʰ/ as they share all but one of their features with /p, t/ and /k/ respectively. The end result is 3/4.

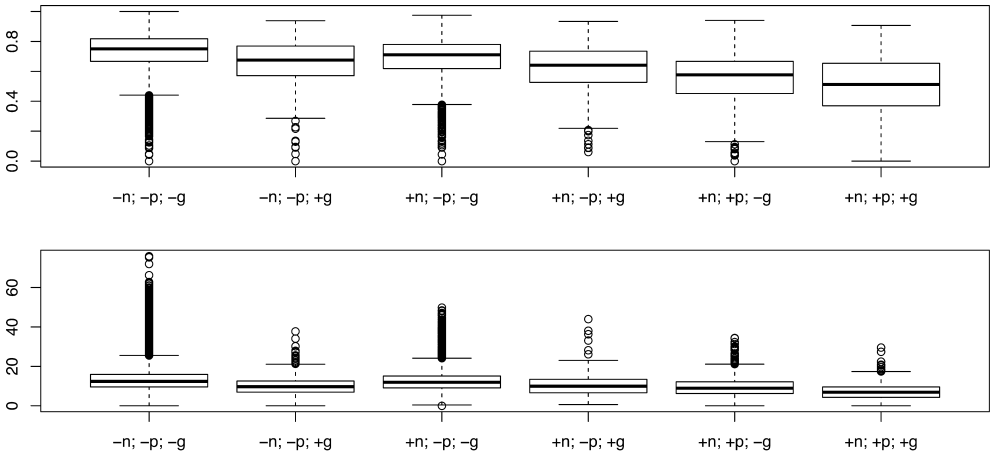

FIGURE 2 Boxplots of Jaccard (top) and CRCJ (bottom) distances between language pairs in different groups

Unlike the Jaccard metric, CRCJ is not normalized (the values do not fall between 0 and 1, but may take arbitrary positive values), which reflects disparities in the sizes of consonantal inventories. In other words, if the Jaccard metric measures the degree of overlap between inventories without taking finer similarities between them into account, CRCJ measures their total difference and pays attention to these similarities.

As a result, as is evident from boxplots of distributions of Jaccard and CRCJ dissimilarity values for the six groups described above shown in Figure 2, CRCJ values, despite being non-normalized, actually have smaller deviations. Also, it may be noted that while the Jaccard distributions are right heavy (they have a long left tail consisting of outlying pairs having low dissimilarity), the CRCJ distributions are left heavy (with long right tails consisting of several language pairs with an unusually big distance between them), which suggests that the Jaccard metric indeed fails to capture similarities between inventories consisting of similar, but not identical segments.

As the distributions of dissimilarities are not normal, we used the Kruskal–Wallis test to check if there are significant differences between groups and then conducted a post-hoc analysis of group differences using the Dunn test.

In the case of Jaccard dissimilarity, the Kruskal–Wallis test showed significant differences between groups at an α-level of 0.05, and the post-hoc analysis confirmed that all between-group differences are significant as well.

Results of tests of between-group differences in CRCJ values are the same, except for the fact that there is no significant difference between groups (+n; −p; −g) and (+n; +p; −g) (p = 0.055). The result is clearly borderline and hinges

on the uncertain merit of 0.05 as the significance level; however, it may indicate that neighboring languages from different phyla display parallel dynamics of segment-inventory development and that this development mostly consists of acquiring distinctions in VOT and additional articulations, in line with the analysis presented by Lindblom & Maddieson (1988).

Based on the results of the post-hoc tests and the differences between median dissimilarity values, we may propose the following hierarchy:

(+n; +p; +g)
< (–n; +p; +g)
< (+n; +p; –g), (+n; –p; –g)
< (–n; +p; –g)
< (–n; –p; –g)

Most importantly, it shows that

1. non-neighboring languages from the same phylum but not the same genus (–n; +p; –g) tend to be more phonologically different between themselves than unrelated neighboring languages (+n; –p; –g); and that
2. neighboring languages from the same genus (+n; +p; +g) tend to be significantly more similar to each other phonologically than non-neighboring languages from the same genus (–n; +p; +g).

Thus, language contact is never insignificant for the development of phonological inventories, at least given the commonly assumed phylogenetic boundaries, and it becomes more significant than phylogeny with the passage of time.[6]

4 Predicting phonological distributions

In this section, I will attempt to predict distributions of individual segments based on areal data. To this end, a new method for quantifying language contact is introduced in § 4.1. In § 4.2, distributions of different consonants are analyzed.

4.1 *Neighbor-pressure metric*

The neighbor graph built according to the procedure described in § 2.2 is used in this study to construct a measure of spatial density for a feature of interest,

6 The answer to the question whether this holds for phyla with possible deeper genetic connections, such as the Indo-European and Uralic families, is a subject for future work.

the neighbor-pressure metric, which provides a quantitative assessment of the extent to which neighbors of a given language are nudging it towards acquiring or preserving the binary feature of interest.

A somewhat simpler way to incorporate language contact into a model is to use pairwise distances (geodesic or 'travel' distances computed according to some procedure) between languages. This approach, although having proved to give good results for certain type of tasks (Jaeger et al. 2011), has several conceptual drawbacks in the context of this study:

1. It does not naturally provide a set of neighbors for a given language. In order to produce a neighbor set for a given language, one has to introduce an arbitrary cut-off distance; the drawbacks of this approach were discussed above. It may be noted that some models do not need a neighbor set, but it is advantageous to have a unified way to incorporate contact data into the calculations, cf. the analyses in §§ 3 and 4.3.
2. It does not correspond to an intuitive notion of 'languages in contact' assumed in the language-contact literature. When regressing on distances between languages, we implicitly introduce a completely new way of thinking about language contact, which, as far as I am aware, has not received any conceptual assessment.
3. It has been reported (Wichmann & Holman 2009; Kalyan & Donohue 2017) that there is a cut-off distance of several thousand kilometers after which languages lose any noticeable influence on one another; this has been corroborated by my own calculations. Therefore, raw distances seem to be a rather noisy predictor variable. A complex transformation can be devised that will make mutual influence decay sublinearly for small distances (there is no real difference if languages are separated by 15 or 20 kilometers) and then make it decay polynomially or even exponentially from some fixed point onwards, but to my knowledge no such function has been proposed to date.

It may also be noted that the regression-on-distance model is tightly coupled with a particular diachronic scenario, that of migration. Jaeger et al. (2011) envisage languages as radiating from Africa and changing along the way. The data-generating process assumed for the data under discussion, on the other hand, views languages as relatively spatially stable during the last millennium, evolving in time and influencing each other in the process—a general case of long-range influence in this model is assumed to be mediated by intervening linguistic communities.

Finally, the nature of the studied variables is significant: Jaeger et al. (2011) analyze distributions of inventory sizes, a variable which can be thought of as decaying with time or travel distance. In this paper, distributions of atomic

units are investigated, and it is evident from the impressionistic analysis of the data that these distributions tend to have relatively sharp boundaries, thus necessitating the addition of a discrete layer to the model.

Therefore, instead of using geographical distances directly, I propose to weight the influence of a language on another language by the distance separating them in the graph—the minimal number of edges that have to be traversed in order to get from one node to another. In order to account for the fact that languages interact mainly with their immediate neighbors, but not to preclude the possibility that languages may be in contact even if they are not connected by an edge in the graph, I propose to assume that the influence of a language on another language decreases exponentially with the length of minimal path between them. The resultant formula is the following:

$$\mathrm{NPM}(l_0, L) = \sum_{l \in L, l \neq l_0} \frac{i_l}{2^{d_l}}$$

where L is the set of all languages in the sample; l_0 is the language for which the density measure is computed; i_l is equal to 0 or 1 according to whether a given language possesses the feature of interest; and d_l is the minimal length of a path between l_0 and l (language of interest) in the neighbor graph.

In other words, in order to compute NPM for a given feature for a given language we add $1/2^{\text{distance in the graph}}$ for each other language that possesses this feature.

4.2 *Modelling consonant distributions with areal data using logistic regression*

Using the NPM metric defined in the previous section, it is possible to investigate difference between consonants as regards the dependence of their distribution on language contact. These differences can be interpreted as the extent to which different segments are prone only to be inherited (vertical transmission) or also to be acquired from neighboring languages (horizontal transmission).

In order to address this question, two nested logistic regression models were fitted for every segment in the data sample that is found in 20+ languages (98 segments).

One model included a factor predictor for phylum and genus combined into a single factor variable (e.g. 'Indo-European-Germanic') with treatment (a.k.a. 'dummy') coding and the other one additionally included NPM. Since simple inclusion of NPM values can lead to possible collinearity between independent variables (as neighboring languages also tend to be related), a simple linear

regression was first fit to the data with phylum+genus as an independent variable and NPM as a response variable. The residuals from this model (i.e. the component in the NPM values not predictable from phylum+genus) were then used in the bigger logistic-regression model.

In order to compute significance values for the effect of NPM, a permutation test was performed following the procedure described by Potter (2005). The rationale for the test, which makes very few assumptions about the distribution behind the data, is the following: under the null hypothesis, the likelihood given the data of the model, which includes NPM as a predictor, should be the same as the likelihood of the smaller model (and their likelihood ratio should be close to 1). Therefore, randomly permuting the NPM residuals we should with equal probability increase or decrease the likelihood of the new model. If we record likelihood ratios for two models for different random permutations of NPM residuals (1000 random permutations were used) and then compute the proportion of the ratios that are smaller than the true ratio, we will obtain the probability that a likelihood ratio as big as that of the true models could have been obtained by chance [cf. also Janssen et al. (2006)].

The *p*-values obtained using this procedure were first corrected for multiple comparison using the Holm method, which showed that there are 10 phonemes significantly dependent on NPM in their distribution (cf. first ten rows in Table 2). It is evident, however, that this method is not powerful enough, in the statistical sense that it too often assumes that the null hypothesis is true.

Indeed, under all the null hypotheses, the distribution of the respective *p*-values is uniform. Therefore, the α-level of 0.05 should give around $98 \times 0.05 \approx 5$ Type I errors (false rejections of the null hypothesis) for 98 tests. Our data in the meantime give 35 *p*-values that are less than 0.05. The probability of obtaining a result of this magnitude by chance can be approximated using Poisson distribution with the mean equal to 5 (by subtracting from 1 the probabilities of obtaining 0, 1, …, 34 spurious results), and this probability is essentially equal to zero.

It must be remembered that familywise error rate methods such as Bonferroni and Holm corrections control the probability of at least one Type I error, which is too strict in our case.[7] As we actually can tolerate some degree of error without prejudicing the validity of the analysis, we instead propose to use the False Discovery Rate method (Benjamini & Hochberg 1995), which controls for the share of falsely rejected null hypotheses. Under an FDR of 0.05 (one in 20 significant *p*-values is likely to be due to chance) we obtain a list of 23 segments

7 'The control of the FWER is important when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is.' (Benjamini & Hochberg 1995: 290)

TABLE 2 Consonants whose distribution is significantly dependent on NPM

Phoneme	*p*-values			Count
	Uncorrected	Holm corrected	FDR corrected	
θ	< 0.001	< 0.001	< 0.001	24
ʈ	< 0.001	< 0.001	< 0.001	124
ʈʰ	< 0.001	< 0.001	< 0.001	80
ɖ	< 0.001	< 0.001	< 0.001	119
ɖʱ	< 0.001	< 0.001	< 0.001	23
ɽ	< 0.001	< 0.001	< 0.001	56
ʂ	< 0.001	< 0.001	< 0.001	82
gʱ	< 0.001	< 0.001	< 0.001	48
ɣ	< 0.001	< 0.001	< 0.001	87
q	< 0.001	< 0.001	< 0.001	85
bʱ	0.001	0.088	0.007	33
sʲ	0.001	0.088	0.007	26
ʐ	0.001	0.088	0.007	46
ɖɽ	0.001	0.088	0.007	36
ɕ	0.002	0.168	0.013	77
v	0.004	0.332	0.025	158
ʦʰ	0.005	0.410	0.029	33
pʔ	0.006	0.486	0.031	20
ð	0.006	0.486	0.031	33
kʔ	0.008	0.632	0.037	25
x	0.008	0.632	0.037	157
ʦ	0.009	0.693	0.040	48
β	0.011	0.836	0.047	33

whose distribution is significantly dependent on the NPM values. The segments and corresponding *p*-values with Holm and FDR corrections are presented in Table 2.

The top of the chart is dominated by retroflex segments (6 out of top 10, 10 out of top 23). Together with voiced aspirated (a.k.a. 'murmured voiced') stops /gʱ, bʱ/ they seem to point to contact processes that took place in and around South Asia (Bashir 2016). Another prominent group are interdental fricatives /θ, ð/, which have a reputation of being 'marked' sounds and are known to be easily lost (Blevins 2006). It is logical therefore that their distribution is largely

shaped by borrowing events (perhaps largely driven by Arabic, a major source of borrowed segments in West Asia).

There is a disparity between the circumspect number of individual phonemes, whose distribution is significantly dependent on NPM, on one hand, and, on the other hand, the evident importance of being neighbors in the language graph for magnitudes of differences between whole inventories highlighted in the previous section. This suggests that neighboring inventories converge by acquiring (or losing) all kinds of segments without evident restrictions (cf. borrowing of clicks in South Africa, of ejectives in the Caucasus, and of pharyngealized segments in West Asia).

4.3 *Non-linear modeling*

The study reported in the previous section showed that for a wide array of consonants, incorporating areal data in the models of their distributions leads to significant increase in explained variance. However, there is also a possibility that the presence of a particular segment in a given inventory may depend on the presence of several segments in neighboring inventories. For example, if several neighboring inventories contain a set of retroflex plosives, but do not contain retroflex affricates, the probability that a retroflex affricate will be present in the inventory in question would still be higher than if no retroflex plosives were present in the neighborhood. Incorporating all interactions of such kind as predictors in a linear model is not feasible.

One way to work around this issue would be to compute local densities of features and feature combinations instead of full segments. However, a series of preliminary models showed that using feature combinations instead of full segments for predicting structures of inventories leads to worse, and not better, fit.

Another option is to use a more complex model, which is able to automatically extract complex dependencies from the data. In addition to estimating the importance of complex dependencies in language contact, these models also make it possible to try to directly incorporate internal inventory data as a predictor. This thus contrasts the contribution of phylogenetic, areal, and structural factors to the distributions of individual consonants.

4.3.1 Data and models

In order to take a fuller account of contact data, the data set used for fitting logistic regression models from the previous section was enriched with exposure sets (binary vectors specifying which segments are present in the inventories of neighboring languages) as well as phylum and genus sets (binary vectors specifying which segments are present in the inventories of languages from the same phylum/genus respectively). Exposure and phylum/genus sets do not

take into account whether a segment is present in one or several neighboring/related inventories: only values of 0 and 1 are used. The analysis based on phylum/genus sets was necessarily based on a smaller dataset because isolates, for which no phylum/genus sets could be constructed, had to be removed.

In order to further incorporate internal data into the analysis, punctured inventory sets were created, where the column corresponding to the segment in question is always zeroed out, but the presence of all other segments is faithfully registered.

The intention is now to compare the prediction accuracy of non-linear models with simpler logistic-regression models. It may be noted that the linear model used for predicting segment distributions in § 4.2 is probably not an optimal one for this task: a regularized regression model, which does not produce unbiased estimators of the parameters, but is more robust with respect to quirks in the distributions of the predictor variables, usually produces better predictions. In order to make the comparison fair, an elastic-net model implemented in the R package glmnet (Simon et al. 2011) was used with all the parameters set to default and the regularization strength was tuned using cross-validation.

A 100-tree random forest classifier with default parameters implemented in the Python package `scikit-learn` (Pedregosa et al. 2011) was used as a non-linear predictor.[8]

4.3.2 Results

In order to test the importance of phylogenetic and areal data for predicting consonant inventories, a series of regularized logistic-regression models was first fit to the data using the following predictors:

1. Basic frequencies (i.e. each segment was predicted for each language if it is present in more than half of all languages)
2. Basic frequencies + phylum + genus
3. Basic frequencies + NPM
4. Basic frequencies + phylum + genus + NPM

8 A deep-neural-network classifier with six fully connected hidden layers consisting of 183, 91, 45, 22, 11, and 5 rectified linear units and a softmax output layer fit using the Python package tensorflow (Abadi et al. 2015) with the default Adagrad optimizer was also fit to the data in order to check if there is complex structure unrecoverable using a random-forest classifier. The performance of the neural net was mostly marginally worse than that of the random forest and is not reported, except for the case when internal inventory data are used as a predictor.

TABLE 3 Performance of regularized logistic-regression models

Model/Metric	Accuracy	ROC AUC
Frequencies	0.837	0.829
Frequencies + phylum + genus	0.84	0.843
Frequencies + NPM	0.857	0.897
Full model	0.862	0.9

TABLE 4 Performance of non-linear classifiers

Model/Metric	Accuracy	ROC AUC
Random forest with phylum sets	0.866	0.886
Random forest with genus sets	0.871	0.897
Random forest with exposure sets	0.884	0.903
Random forest with internal data	0.91	0.979

Accuracy scores and the AUC-ROC metric for all four models obtained by 10-fold cross-validation are reported in Table 3.

The results indicate that NPM on its own is a better predictor for distributions of Eurasian consonants than phylogenetic data, but that the best predictive performance is achieved when they are combined.

Performance of these models was then compared with performance of a random-forest classifier, which, in addition to phylogenetic information, used exposure/phylum/genus sets or internal inventory data. Both classifiers were trained on random 90% of the data with 10% of data points used for validation. The procedure was repeated several times with nearly identical results, and only the results of the last run are reported in Table 4.

A random forest with any of exposure/phylum/genus sets outperforms logistic-regression models, which indicates that phylogeny and NPM used as linear predictors do not adequately explain distributions of segments. A random forest working from exposure sets furthermore clearly beats the model based on phylum sets and performs better than the one using genus sets. As in the case of using exposure/phylum/genus sets we cannot effectively decorrelate the predictors, it is impossible to contrast the effect of language contact and phylogeny directly, but it seems that exposure sets are a better predictor in general (it also must be remembered that the exposure-set classifier also covers lan-

guage isolates, for which no predictions can be made based on phylogeny). The incorporation of both internal inventory information and exposure sets as predictors counterintuitively lead to degraded performance, presumably because of a lower signal-to-noise ratio in the resulting data.

Interestingly, but not entirely surprisingly, internal inventory data provide a more solid basis for predicting presence or absence of a particular consonant. That means that taking an inventory table and covering one of the cells we may very accurately predict whether it is filled or not knowing only basic frequencies (even if we drop phylum and genus as predictors, accuracy stays above 0.91 for a random forest and above 0.925 for a neural network). This indicates that there may exist universal tendencies in inventory structures, akin to those explored in Dunbar & Dupoux (2016), which are better captured by the neural network model than by the random forest.

The overall difference in performance between logistic-regression models and non-linear classifiers having access to inventories of neighboring/related languages is noticeable (the accuracy rises from ≈ 0.86 to ≈ 0.9) but is actually hardly crucial for typological studies mostly interested in overall trends. Therefore it may be concluded that NPM as a proxy measure of language contact is an adequate instrument for controlling for language contact in typological studies and can be used in this capacity instead of more coarse-grained macroarea-based variables.

Conclusion

In this paper, the following findings were presented:

1. Based on full-inventory differences defined using the Jaccard dissimilarity and the novel CRCJ (Closest-Relative Cumulative Jaccard Dissimilarity) metric, it was demonstrated that language contact consistently influences inventory structures: neighboring languages from the same phylum and even from the same genus are significantly more similar phonologically than non-neighboring ones. Moreover, according to the CRCJ metric, neighboring languages from the same phylum, but not the same genus, are not significantly more phonologically similar to each other than neighboring languages from different phyla (p-value for the difference between medians is equal to 0.055; more data is needed to clarify this issue).
2. A novel measure of exposure of a language to a particular segment—the Neighbor Pressure Metric (NPM)—was shown to be a significant predictor for the geographical distribution of 23 out of 98 segments found in 20+

Eurasian languages, a good share of them connected to the South Asian linguistic area. It was also shown that on its own NPM is at least as good a predictor for the presence of a particular consonant in a language as phylogenetic data in the shape of phylum and genus labels.

3. There exist complex statistical dependencies between the inventory of a given language and full inventories of its neighboring languages. Nonlinear classifiers incorporating unions of inventories of neighboring or related languages as predictors perform significantly better than logistic-regression models based only on single-segment information, and inventories of neighboring languages tend to be a slightly better predictor than inventories of related languages.

Together these findings highlight the necessity for developing nuanced models of phonological-inventory development, which must take into account not only phylogenetic and socio-geographical factors, but also language contact. Conversely, they show that for typological studies of consonant-inventory structures it is imperative to incorporate language contact as a control variable and that for binary and possibly multivariate features it is possible to do so by using NPM as defined on a suitable neighbor graph.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu & Xiaoqiang Zheng. 2015. *TensorFlow: Large-scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. https://www.tensorflow.org/.

Atkinson, Q.D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027). 346–349. doi: 10.1126/science.1199295.

Bashir, Elena. 2016. Contact and convergence. In Hans Henrich Hock & Elena Bashir (eds.), *The Languages and Linguistics of South Asia: A Comprehensive Guide*, 241–374. Berlin: de Gruyter Mouton.

Benjamini, Yoav & Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* (*Methodological*) 50(1). 289–300.

Bickel, Balthasar. 2013. Distributional biases in language families. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language Typology and Historical Contingency: In Honor of Johanna Nichols*, 415–444. Amsterdam: John Benjamins Publishing Company.

Bickel, Balthasar & Johanna Nichols. 2006. Oceania, the Pacific Rim, and the theory of linguistic areas. *Proc. Berkeley Linguistics Society* 32. 3–15.

Blevins, Juliette. 2006. New perspectives on English sound patterns: "Natural" and "unnatural" in Evolutionary Phonology. *Journal of English Linguistics* 34(1). 6–25.

Campbell, Lyle. 2006. Areal linguistics: A closer scrutiny. In Yaron Matras, April McMahon & Nigel Vincent (eds.), *Linguistic Areas: Convergence in Historical and Typological Perspective*, 1–31. Basingstoke: Palgrave Macmillan.

Campbell, Lyle. 2017. Why is it so hard to define a linguistic area? In Raymond Hickey (ed.), *The Cambridge Handbook of Areal Linguistics*, 19–39. Cambridge: Cambridge University Press.

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244.

Clements, G.N. 2003. Feature economy in sound systems. *Phonology* 20(03). 287–333.

Cotterell, Ryan & Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. CoRR abs/1705.01684. http://arxiv.org/abs/1705.01684.

Daumé III, Hal. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 593–601*. Stroudsburg, PA Association for Computational Linguistics.

Delaunay, Boris. 1934. Sur la sphère vide. Á la mémoire de Georges Voronoï. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et naturelles* 6. 793–800.

Donohue, Mark & Johanna Nichols. 2011. Does phoneme inventory size correlate with population size? *Linguistic Typology* 15(2). 161–170.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257292.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Dunbar, Ewan & Emmanuel Dupoux. 2016. Geometric constraints on human speech sound inventories. *Frontiers in Psychology* 7. doi: 10.3389/fpsyg.2016.01061.

Ember, Carol R. & Melvin Ember. 2007. Climate, econiche, and sexuality: Influences on sonority in language. *American Anthropologist* 109(1). 180–185.

Emeneau, Murray B. 1956. India as a Linguistic Area. *Language* 32(1). 3–16.

Enfield, Nick J. 2005. Areal linguistics and mainland Southeast Asia. *Annual Review of Anthropology* 34. 181–206.

Everett, Caleb. 2013. Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PLoS ONE* 8(6). e65275.

Fortescue, Michael. 1998. *Language Relations Across Bering Strait: Reappraising the Archaeological and Linguistic Evidence*. London: Cassell.

Fought, John G., Robert L. Munroe, Carmen R. Fought & Erin M. Good. 2004. Sonority and climate in a world sample of languages: Findings and prospects. *Cross-Cultural Research* 38(1). 27–51.

Friedman, Victor A. 2000. After 170 years of Balkan Linguistics: Whither the Millennium? *Mediterranean Language Review* 12. 1–15.

Gordon, Matthew. 2016. *Phonological Typology*. Oxford: Oxford University Press.

Güldemann, Tom. 2008. The Macro-Sudan belt: Towards identifying a linguistic area in northern sub-Saharan Africa. In Bernd Heine & Derek Nurse (eds.), *A Linguistic Geography of Africa*, 151–185. Cambridge: Cambridge University Press.

Hay, Jennifer. & Laurie Bauer. 2007. Phoneme inventory size and population size. *Language* 83(2). 388–400.

Hockett, Charles F. 1955. *A Manual of Phonology*. Baltimore: Waverley Press.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–320.

Janssen, Dirk P., Balthasar Bickel & Fernando Zúñiga. 2006. Randomization tests in language typology. *Linguistic Typology* 10(3). 419–440.

Kalyan, Siva & Mark Donohue. 2017. Mapping hotspots of morphosyntactic and phonological diversity. A paper presented at the 12th Conference of the Association for Linguistic Typology, Canberra, Australia.

Ladd, D. Robert, Séan G. Roberts & Dan Dediu. 2015. Correlational studies in typological and historical linguistics. *Annual Review of Linguistics* 1(1). 221–241. doi:10.1146/annurev-linguist-030514-124819.

Liljencrants, Johan & Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48(4). 839–862.

Lindblom, Björn & Ian Maddieson. 1988. Phonetic universals in consonant systems. In Larry M. Hyman & Charles N. Li (eds.), *Language, Speech, and Mind: Studies in Honour of Victoria A. Fromkin*, 62–78. London: Routledge: published in the USA in association with Routledge, Chapman and Hall.

Maddieson, Ian. 1991. Testing the universality of phonological generalizations with a phonetically specified segment database: Results and limitations. *Phonetica* 48(2–4). 193–206.

Maddieson, Ian. 1984. *Patterns of sounds*. Cambridge: Cambridge University Press.

Maddieson, Ian & Kristin Precoda. 1992. *UPSID and PHONEME* (*version 1.1*). Los Angeles: University of California at Los Angeles.

Michael, Lev, Will Chang & Tammy Stark. 2014. Exploring phonological areality in

the Circum-Andean region using a naive Bayes classifier. *Language Dynamics and Change* 4(1). 27–86.

Michael, Lev, Tammy Stark, Emily Clem & Will Chang (eds.). 2015. *South American Phonological Inventory Database v1.1.4*. Berkeley: University of California. http://linguistics.berkeley.edu/saphon/

Mielke, Jeff. 2008. *The Emergence of Distinctive Features*. Oxford: Oxford University Press.

Moran, Steven, Daniel McCloy & Richard Wright. 2012. Revisiting population size vs. phoneme inventory size. *Language* 88(4). 877–893.

Moran, Steven, Daniel McCloy & Richard Wright (eds.). 2014. *PHOIBLE online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://phoible.org/.

Murawaki, Yugo & Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution* 3(1). 13–25.

Nettle, D. 2012. Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1597). 1829–1836.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nikolaev, Dmitry, Andrey Nikulin & Anton Kukhto (eds.). 2015. *The database of Eurasian phonological inventories* (beta version). http://eurasianphonology.info/.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825– 2830.

Potter, Douglas M. 2005. A permutation test for inference in logistic regression with small-and moderate-sized data sets. *Statistics in Medicine* 24(5). 693–708.

da Silva, Sara Graça & Jamshid J. Tehrani. 2016. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science* 3(1). 150645.

Simon, Noah, Jerome Friedman, Trevor Hastie & Rob Tibshirani. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* 39(5). 1–13. http://www.jstatsoft.org/v39/i05/.

Stevens, Kenneth N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17(1). 3–45.

Thomason, Sarah Grey. 2000. Linguistic areas and language history. *Studies in Slavic and General Linguistics* 28. 311–327.

Towner, Mary C., Mark N. Grote, Jay Venti & Monique Borgerhoff Mulder. 2012. Cultural macroevolution on neighbor graphs. *Human Nature* 23(3). 283–305.

Trubetzkoy, Nikolai S. 1939. *Grundzüge der Phonologie*. Prague.

Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Wichmann, Søren. 2017. Genealogical classification in historical linguistics. In Mark Aronoff (ed.), *Oxford Research Encyclopedia of Linguistics*, http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-78.

Wichmann, Søren, Taraka Rama & Eric W. Holman. 2011. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15(2). 177–197.

Wichmann, Søren & Eric W. Holman. 2009. *Temporal Stability of Linguistic Typological Features*. Munich: Lincom Europa.

Yamauchi, Kenji & Yugo Murawaki. 2016. Contrasting vertical and horizontal transmission of typological features. In *Proc. of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 836–846.