Ian Joo* and Yu-Yin Hsu

Phonological distances between Eurasian lects measured via Phonotacticon 1.0 reveal areal patterns

https://doi.org/10.1515/ling-2024-0157
Received August 17, 2024; accepted September 6, 2025; published online April 29, 2026

Abstract: The present study measures the phonological distances between 335 spoken lects of Eurasia using Phonotacticon 1.0, a cross-linguistic database, and explores whether phonologically similar lects form areal patterns within Eurasia. Results indicate that phonological clusters tend to form geographical clusters, which divides Eurasia horizontally into eastern (East/Southeast Asia), central (South/Central/North/West Asia), and western (Europe) regions. The convergence patterns in the phonological domain overlap with morphosyntactic convergence patterns measured via Grambank to some degree, but not entirely, suggesting the domain-specific nature of areal convergence. Finally, comparison with the genealogical distances between the sample lects shows that morphosyntactic distances, but not phonological distances, are significantly correlated with the number of shared genealogical layers, implying that morphosyntax is more conservative to genealogical heritage compared to phonology.

Keywords: Eurasia; phonotactics; typology

1 Introduction

How close is English phonology to French phonology? Is the phonological distance between English and French closer than the phonological distance between English and Mandarin Chinese? Surely, there are many phonological features that English and French have in common but not shared by Mandarin, such as complex onsets and obstruent codas. Mandarin also has phonological characteristics that distinguish it from both English and French, such as lexical tones and retroflex

***Corresponding author: Ian Joo [tɛʰù.i.an]**, Otaru University of Commerce, 3 Chome-5-21 Midori, Otaru, Hokkaido, 047-0034 Japan, E-mail: joo@res.otaru-uc.ac.jp. https://orcid.org/0000-0001-7678-1227
Yu-Yin Hsu [ɛy^{21}.jou^{51}.jin^{214}], The Hong Kong Polytechnic University, 11 Yuk Choi Rd, Hung Hom, Hong Kong, E-mail: yu-yin.hsu@polyu.edu.hk. https://orcid.org/0000-0003-4087-4995

Open Access. © 2026 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

consonants. But how can we quantify these featural differences to compare one distance to another?

The goal of this paper is to quantify the phonological distance between different Eurasian lects (linguistic varieties, commonly referred to as LANGUAGES or DIALECTS) to compare the distance between a pair of lects to the distance between another pair. The ultimate goal is to detect phonologically similar lects within Eurasia to see whether they form areal patterns. In order to attain this goal, we employ Phonotacticon 1.0 (Joo and Hsu 2025), a cross-linguistic database of the basic phonotactic information of 516 spoken lects of Eurasia.

2 Background

The term PHONOLOGICAL DISTANCE is ambiguous and may refer to two different concepts. One meaning is the distance between the forms of two phonological sequences (lect-internally or cross-linguistically), such as measuring whether /mæn/ is closer to /pæn/ than it is to /kæn/. As an example, Do and Lai (2021) provide a model of measuring the distance between two phoneme sequences, combining segmental and suprasegmental features. We name this type of phonological distance INTERSEQUENTIAL PHONOLOGICAL DISTANCE.

The second meaning is the distance between the phonological structures of two lects. How close is English phonology to Turkish phonology, in terms of phonological inventory, phonotactic constraints, or segmental frequency? And is the distance closer than the distance between English phonology and Japanese phonology? In contrast to the intersequential phonological distance, we call this type of phonological distance INTERSTRUCTURAL PHONOLOGICAL DISTANCE.

Various papers have measured interstructural phonological distances in different ways: based on distinctive features (Afendras 1970; Avram 1964; Kučera and Monroe 1968; Nikolaev 2019; Postovalova 1966), segmental frequency (Tambovtsev 2001), text or audio corpora (Eden 2018), phonotactic constraints (Macklin-Cordes et al. 2021), or acoustic qualities (Harnud and Zhou 2021). The diversity of the methodologies suggests that there is no one correct solution to the problem of quantifying phonological distance, but many possible ways.

One of the areas not covered sufficiently by previous work, however, is comparing multisegmental sequences (for example, the English complex onset /spl-/ to a complex onset of another lect), as opposed to comparing singleton segments (for example the English phoneme /s/ to a phoneme of another lect). In this paper, we will employ Phonotacticon (Joo and Hsu 2025), which contains the sequential information of the onset, nucleus, and coda position of each spoken lect, to fill in this gap.

3 Distance measuring

In this section, we will explain our phonological distance measuring method as follows. First, we present the two databases we use for the distance measuring, namely Phonotacticon 1.0 (Joo and Hsu 2025) and PanPhon (Mortensen et al. 2016) (Section 3.1). Before using these databases, we tidy up the Phonotacticon data by modifying or excluding sample lects with underspecified information, i.e. lects that do not contain sufficient information on what sequences are allowed in onset/nucleus/coda positions (Section 3.2). Next, we measure the distance between each pair of sequences, e.g. the distance between /p/ and /pl/ or that between /pl/ and /spl/ (Section 3.3). Based on these sequence-to-sequence comparisons, we calculate the distance between the onset/nucleus/coda inventory of each lect to the corresponding inventory of another, i.e. the onset inventory (= the set of sequences permitted in the onset position) of a lect versus the onset inventory of another, and so on (Section 3.4). Next, we calculate the tonal distance between each pair of lects based on the number of tones (Section 3.5). Finally, based on the onset/nucleus/coda/tone distances, we calculate the overall phonological distances between each pair of lects (Section 3.6).

3.1 Databases

We measure the phonological distance between Eurasian lects using Phonotacticon 1.0, which is available at zenodo.org/records/10623743. It consists of each of the 516 Eurasian spoken lects' segmental sequences in the onset, nucleus, or coda position. For example, /spl/ is an onset sequence of English, composed of three segments: /s/, /p/, and /l/. Note that onsets, nuclei, or codas that consist of only one segment are also referred to as "sequences" here. In addition, the tones of each lect (or lack thereof) are also recorded in Phonotacticon. Finally, a lect may contain an additional explanatory note.

As an example, Table 1 shows the entry of Jiongnai Bunu `[jion1236]` (Hmong-Mien, Mao and Li 2002). <#> represents an empty onset or an empty coda. For further details of the database, see Joo and Hsu (2025).

Another database we employ for our analysis is PanPhon (Mortensen et al. 2016), which contains an exhaustive list of IPA segments and the featural values of each segment. We have slightly modified PanPhon to render it more compatible with our analysis (the modified version is available at zenodo.org/records/10623743).

Based on the featural vectors and the number of tones per lect, we can calculate how phonologically similar two Eurasian lects are to each other. The notion of phonological similarity here, however, is limited to what sequences are permitted in

Table 1: Phonotacticon 1.0 entry of Jiongai Bunu.

Section	Content
Phoneme	p pʰ ⁿb ⁿp m m̥ f v v̥ w w̥ θ ð t tʰ ⁿd ⁿt n n̥ l l̥ ʧ ʧʰ ⁿd͡ʒ ⁿʧ ʃ ɲ ɲ̥ ɮ ʑ̥ k kʰ ᵑg ᵑk ŋ ŋ̥ x i e a ɔ o u ə j
Tone	44 33 53 31 35 22 43 32
Onset	p pʰ ⁿb ⁿp m m̥ f v v̥ w w̥ θ ð t tʰ ⁿd ⁿt n n̥ l l̥ ʧ ʧʰ ⁿd͡ʒ ⁿʧ ʃ ɲ ɲ̥ ɮ ʑ̥ k kʰ ᵑg ᵑk ŋ ŋ̥ x pl pʰl ⁿpl kl kʰl ᵑgl pj ⁿbj mj θj ðj tj tʰj ⁿdj nj ɲj lj ʎ ʧ ʃ kj kʰj ᵑgj ᵑkj pw pʰw ⁿbw mw θw ðw tw tʰw ⁿdw ⁿtw lw ʧw ⁿd͡ʒw ʃw ɲw kw kʰw ᵑgw klw
Nucleus	ŋ i e a au ɔ ɔu o u ui ə
Coda	# n ŋ
Note	<v̥ ʑ̥> transcribed as /v̥ ʑ̥/.

the onset/nucleus/coda positions and the number of tones due to the nature of the database. There is certainly more to a lect's phonology than just the number of its tones and the featural and phonotactic values of its segments, such as allophonic variations, morphophonemic rules, stress patterns, and phonotactic restrictions beyond syllabic boundaries.

Yet, we argue that the distance measured via Phonotacticon is rightfully called "phonological distance" (and not "phonotactic distance") because we can retrieve from the database most of the phonological parameters that are typically used by typologists when making phonological comparisons. To illustrate this point, we refer to another database as an example, the World Atlas of Language Structures (WALS) Online (Dryer and Haspelmath 2013), a database containing the 192 typological features and the featural values of thousands of lects worldwide. Among the 192 features, twenty are phonological, and of the twenty phonological features, only four features would be completely irretrievable from Phonotacticon: FIXED STRESS LOCATIONS, WEIGHT-SENSITIVE STRESS, WEIGHT FACTORS IN WEIGHT-SENSITIVE STRESS SYSTEMS, and RHYTHM TYPES. Other than these suprasegmental features, all other phonological features, such as LATERAL CONSONANTS and VOWEL NASALIZATION, are at least partially or indirectly retrievable from Phonotacticon. Thus, in comparison with a well-cited cross-linguistic database such as WALS, the phonological information we can gain from Phonotacticon is not exceedingly imbalanced (at least for the purpose of typological comparison), which justifies the Phonotacticon-based distance being named "phonological distance".

3.2 Underspecified sample lects

In Phonotacticon, some lects are described with underspecified segments (such as <P> for plosives) or sequences with brackets (such as [p t k][l r] for any sequence with

/p/, /t/, or /k/ as the first segment and /l/ or /r/ as the second segment). For computational purposes, we converted these underspecified segments and sequences into specific segments and sequences. For example, a lect that has /P/ as a possible onset goes through the process of converting /P/ into all plosive phonemes it has. A lect that has [p t k][l r] as possible onset sequences goes through converting [p t k][l r] into all the logically possible combinations, i.e. /pl pr tl tr kl kr/.

We also excluded sample lects that have sequences including two or more consecutive <C> symbols (representing "consonant"), such as <CC> or <CCC>. This is because such transcriptions would lead to too many possible sequences. For example, one of the English `[stan1293]` coda sequences as coded in the database is /CCCs/ (Gut 2009). Converting this into any three English consonants followed by /s/ would generate too many typologically improbable clusters, such as /ssss/ or /ltms/. Thus, lects whose sequences include consecutive C's were excluded.

Likewise, lects that have bracketed segments with too many members were excluded. For example, one of the possible onsets of Czech `[czec1258]` as coded in the database is [p b f v m t d s z n c ɟ ʃ ʒ ɲ k g x h l r r̝ j][p b t d g f v s z ʃ ʒ x j r r̝ l m n ɲ] (Bičan 2011). This means any biconsonantal sequence whose first member is any one of the 23 segments within the first brackets followed by any one of the 19 segments within the second brackets. As this would generate 437 logically possible sequences, including many onset sequences that do not exist in Czech (such as /pb/, /bt/, or /fm/), it would be overly problematic. All sequences involving ten or more segments within brackets followed by ten or more segments within brackets were excluded.

Approximately a fourth of the sample lects were excluded for having either consecutive C's or a sequence of ten or more bracketed segments. Moreover, lects with missing information in any one of the onset/nucleus/coda slots were excluded. This leaves us with a subset of 335 sample lects out of 516. As a reviewer pointed out, this exclusion biases the sample against lects with more complex syllable structures, as those are more likely to be described by consecutive C's compared to lects with simpler syllable structures.

3.3 Measuring the distance between sequences

In this section, we will show how we measure the distance between two sequences, e.g. between /pl/ and /spl/.

In order to measure the distance between sequences, it is necessary to measure the distance between segments. In measuring the segmental distance, we employ Saporta's (1955) method, henceforth referred to as the SAPORTA DISTANCE. The Saporta distance is the Manhattan distance between the two vectors of featural values, each of which may be of 1 (positive), −1 (negative), or 0 (absent). Saporta (1955)

distinguishes a (full) contrast between a positive featural value and a negative one, which he weighs as two units of difference, from a "semi-contrast" between a positive or negative featural value and the absence of a featural value, which he weighs as one unit of difference. Here, we adopt his view that a full contrast between two opposing values can be weighed heavier than a contrast between a value and the absence of it, since the former is the distance between two opposing values whereas the latter is the distance between one value and the absence of it. To make an analogy to visual perception, the contrast between two complementary chromatic colors (color pairs that have opposing hues, such as blue vs. yellow or red vs. cyan) is perceptually more salient than the contrast between a chromatic color and an achromatic color (i.e. black, white, and the shades in between), which lacks hues (analogous to featural values in phonology).

As an example, Table 2 shows the featural values of /t/ and /p/. The gap column shows the difference between each of /t/'s featural value and the corresponding

Table 2: The Saporta distance between /t/ and /p/.

Feature	t	p	GAP
Syllabic	−1	−1	0
Sonorant	−1	−1	0
Consonantal	1	1	0
Continuous	−1	−1	0
Delayed release	−1	−1	0
Lateral	−1	−1	0
Nasal	−1	−1	0
Strident	0	0	0
Voiced	−1	−1	0
Spread glottis	−1	−1	0
Constricted glottis	−1	−1	0
Anterior	1	1	0
Coronal	1	−1	2
Distributed	−1	0	1
Labial	−1	1	2
High	−1	−1	0
Low	−1	−1	0
Back	−1	−1	0
Round	−1	−1	0
Velaric	−1	−1	0
Tense	0	0	0
Long	−1	−1	0
SUM			5

featural values of /p/. For instance, /t/'s [syllabic] feature is −1 and /p/'s is also −1, their gap being 0, whereas the [labial] feature of /t/ is −1 and that of /p/ is 1, the gap being 2. The sum of these gaps is 5, which is the Saporta distance between /t/ and /p/.

We will apply the Saporta distance method to measure the distance between sequences. For example, in order to compare the distance between /pl/ and /spl/, we calculate the distance between all the logically possible mappings between the two sequences within six segment slots, as shown in Table 3. The number of segment slots was chosen as six as it is the maximal number of sequences among the sample lects. The distance between the two sequences in a given meaning is defined as the sum of the distances between each mapped pair, including the mappings where one member is an empty slot. An empty slot is considered to be a segment that has the value 0 for all phonological features. Among the possible mappings shown in Table 3 (which does not include duplicate mappings with the same distance value), the third mapping yields the minimal distance between /spl/ and /pl/, as the sum of the distance between /s/ and zero values (20), between /p/ and /p/ (0), and between /l/ and /l/ (0). The distance between /spl/ and /pl/ is thus 20 + 0 + 0 = 20.[1] Note that, since the segment slots are limited to six, this model does not yield all logically possible mappings for longer sequences: comparing a sequence of six segments to another sequence with six segments, for example, will yield only one mapping.

As an example, Table 4a shows the twenty sequences that are the most similar to /pl/ and Table 4b those to /ia/. Note that [a], [ă], and [æ] are not featurally distinct in PanPhon, as they are all low front unrounded vowels. The distance between /ia/, /iă/, and /iæ/ is thus zero.

Table 3: Five possible mappings between /spl/ and /pl/.

	1	2	3	4	5	6	Distance
Mapping 1		s	p	l			64
	p	l					
Mapping 2	s	p	l				40
	p	l					
Mapping 3	**s**	**p**	**l**				**20**
		p	**l**				
Mapping 4	s	p	l				72
			p	l			
Mapping 5	s	p	l				98
				p	l		

1 We thank Huisu Yun for providing us with this idea.

Table 4: Sequences most similar to /pl/ and /ia/.

Sequence	Distance
(a) /pl/	
pl	0
pˣl	1
bl	2
pʰl	2
p̥l	2
pʼl	2
ⁿpl	2
pˀl	2
plʲ	2
fl	3
p͡fl	3
plˠ	4
m̥l	4
b̥l	4
b̤l	4
pʰl̥	4
pz	4
ⁿbl	4
bˀl	4
ɓl	4
(b) /ia/	
ia	0
iă	0
iæ	0
ie	2
ea	2
iã	2
iĕ	2
iæ̃	2
ɪa	2
ia:	2
i:a	2
ĩa	2
ɨa	2
iə̯	2
eæ	2
iæ:	2
i:æ	2
iɑ	3
iɛ	4
i:e	4

3.4 Measuring the inventory distance between lects

In this section, we will show how we measure the distance between two lects in terms of onset, nucleus, and coda inventories.

We calculate the distance between two lects within the same category (onset, nucleus, or coda). The distance between the onset/nucleus/coda sequences of two lects is defined as follows. Let M^A and M^B be the matrices representing the phonological feature values of the onset/nucleus/coda sequences of lect A and lect B, respectively. Let MD be the distance matrix between M^A and M^B. The distance between the onset/nucleus/coda forms of A and B is the average value of the minimum values of rows or columns of MD, whichever is higher.

As an example, suppose that A allows three onset sequences, /p m t/, and B two onset sequences, /p t/. (Although /p m t/ are singleton segments, the term SEQUENCE is meant to include singleton segments as "sequences" with only one segment.) The comparison between A and B is shown in the Table 5. The last column and the last row shows the minimum value of each row and each column, respectively. The average value of the minimum column (the comparison from lect 1 to lect 2) is (0 + 6 + 0)/3 = 2, whereas the average value of the minimum row (the comparison from B to A) is 0. The bigger value of these two is selected. Thus, the onset distance between A and B is 2.

Using this formula, we calculate the distance of onset, nucleus, and coda of each pair of lects.

3.5 Measuring the distance between tones

Next, we calculate the distance between the tones of each pair of lects. Tones are suprasegmental distinctions of lexemes, mainly by pitch, but often in combination with other acoustic cues, such as length, creakiness, and breathiness. In Eurasia, tonal lects are mostly concentrated in Mainland Southeast Asia, although they also

Table 5: Comparison between lect A with the onset sequences /p m t/ and lect B with the onset sequences /p t/.

A/B	p	t	Min
p	0	5	0
m	6	11	6
t	5	0	0
Min	0	0	

exist in other areas, such as Swedish [swed1254] (Riad 2013) in Europe and Eastern Panjabi [panj1256] (Bhatia 1993) in South Asia.

The distance between tonality is defined as the Canberra distance between the numbers of tonemes of two lects. Let T1 be the number of tonemes lect 1 has and T2 the number of tonemes lect 2 has. The distance betwen lect 1 and lect 2 is:

$$\frac{|T1 - T2|}{T1 + T2} \tag{1}$$

For example, Burmese [nucl1310] (Jenny and Hnin Tun 2016) has three tonemes, whereas Yue Chinese [yuec1235] (Bauer and Benedict 1997) has six. The tonal distance between two lects is thus:

$$\frac{|3 - 6|}{3 + 6} = \frac{1}{3} \tag{2}$$

If both lects have 0 tonemes, then the distance between the two lects is 0.

3.6 Measuring the overall distance

We then calculate the overall distance, which is the Euclidean distance between each pair of lects based on their four normalized distances (onset, nucleus, coda, and tone).

Admittedly, assigning equal weight to the four types of distances is a rather simplistic approach. As a reviewer of a previous version of this paper pointed out, given the wider articulatory variance of consonants compared to that of vowels, the distance of onsets and codas, which mainly consist of consonants, may merit more weight than the distance of nuclei, which mainly consist of vowels. Additionally, weighing the tonal distance equally to the three segmental distances may also be a problem, since unlike the segmental distances, the tonal distances are not normally distributed. As there are more atonal lects than tonal lects, the tonal distances are unevenly distributed to the two extremes of 0 (two atonal lects or two tonal lects with the same number of tones) and of 1 (a tonal lect versus an atonal lect). Paradoxically, weighing tones equally as segmental categories lects may result in a wider distance between tonal and atonal lects compared to the distance between lects that may not differ in tonality but may differ significantly in terms of segmental phonotactics. We chose this method of equal weighing regardless, given that it is difficult to determine exactly how much weight should be assigned to the tonal difference compared to segmental differences and naively assuming equal weight seemed to be the optimal solution. A more sophisticated weighing of the four types of distances that does justice to the different levels of variance and

Table 6: Overall distance between Thai and Northeastern Thai.

	Thai	Northeastern Thai	Distance
Onset	b d p t c k ʔ pʰ tʰ cʰ kʰ f s h m n ŋ l r w j pr tr kr kw pl kl pʰr kʰr kʰw pʰl kʰl	p t k ʔ pʰ tʰ kʰ b d c͡ɕ m n ɲ ŋ f s h l w j	1.13
Nucleus	i ɯ u e ə o ɛ a ɔ iː ɯː uː eː əː oː ɛː aː ɔː ia ɯa ua	i iː ɨ ɨː u uː e eː ə əː o oː ɛ ɛː a aː ɔ ɔː iə ɨə uə	0.28
Coda	# p t k ʔ m n ŋ w j	# p t k ʔ m n ŋ w j	0.00
Tone	H M L R F	M L H F R MR	0.19
Overall			1.18

distribution of the four parameters is warranted in the future analyses of phonological distance using Phonotacticon.

The full list of phonological distances measured using this methodology is available at `ianjoo.github.io/PhonoDist.csv`. The R Markdown script used for the calculation can be downloaded from `github.com/ianjoo/PhonotacticonLinguistics`.

As an example of the overall distance between two closely related lects, Table 6 shows the calculation of the overall distance between (Central) Thai `[thai1261]` (Iwasaki and Ingkaphirom 2005) and Northeastern Thai `[nort2741]` (Kongsin 1979). The onset/nucleus/coda/tone distances refer to the normalized distance values minus the minimal value (rounded to the second decimal), whereas the overall distance refers to the Euclidean distance of the four distances. We see that the codas of the two lects are identical, the number of tones differ only by one, the nuclei are largely similar, and the biggest difference lies in the onsets, Thai permitting complex onsets whereas Northeastern Thai does not.[2]

To put it into perspective, Table 7 shows the overall distance calculation between Thai and Dutch `[dutc1256]` (Booij 1999). In all four dimensions, the distance is greater. The biggest difference is the number of tones, as Dutch completely lacks tones. The difference between coda distances of the two pairs is also significant, as Dutch allows complex codas, unlike Thai or Northeastern Thai. Onset and nucleus distances between Thai and Dutch also differ from those between Thai and Northeastern Thai, albeit to a lesser degree.

As an example of how the measured distances reflect areal convergence patterns, Table 8 shows the twenty sample lects closest to Hindi `[hind1269]` (Kachru 2006). We see that, unsurprisingly, the majority of the lects are spoken in South Asia.

2 Note, however, that complex onsets are frequently simplified in Central Thai as well, depending on the sociocultural context (Beebe 1975).

Table 7: Overall distance between Thai and Dutch.

	Thai	Dutch	Distance
Onset	b d p t c k ʔ pʰ tʰ cʰ kʰ f s h m n ŋ l r w j pr tr kr kw pl kl pʰr kʰr kʰw pʰl kʰl	# p b t d k g f v s z x ɣ h m n ŋ l r ʊ j pj tj kj tw dw sw zw fj sj pl bl pr br tr dr fl vl fr vr sl xl xr ɣl ɣr sm sn kn pn fn ɣn sp st sk sf sx spr str skr spl skl sxr	1.23
Nucleus	i ɯ u e ə o ɛ a ɔ iː ɯː uː eː əː oː ɛː aː ɔː ia ɯa ua	ɪ ɛ ɔ ɤ ɑ i y u e ø o a ə ɛi œy ɔu	0.67
Coda	# p t k ʔ m n ŋ w j	# p b t d k g f v s z x ɣ m n ŋ l r ʊ j lm rm rn lp lf lv lk lɣ rp rf rv rk rx rɣ sp st sk	1.24
Tone	H M L R F	–	2.11
Overall			2.82

Table 8: The twenty lects closest to Hindi.

Family	Lect	Distance
Dravidian	Jennu Kurumba	0.55
Indo-European	Kotia-Adivasi Oriya-Desiya	0.62
Indo-European	Sindhi	0.67
Austroasiatic	Korku	0.81
Indo-European	Lambadi	0.89
Indo-European	Konkan Marathi	0.92
Nihali	Nihali	0.97
Dravidian	Korra Koraga	1.03
Indo-European	Saurashtra	1.04
Dravidian	Sholaga	1.12
Dravidian	Kui (India)	1.14
Dravidian	Muduga	1.18
Indo-European	Kashmiri	1.22
Sino-Tibetan	Duhumbi	1.23
Dravidian	Kodava	1.25
Indo-European	Halbi	1.26
Indo-European	Nuristani Kalasha	1.27
Indo-European	Vaagri Booli	1.27
Uralic	Pite Saami	1.28
Dravidian	Waddar	1.30

(Urdu [urdu1245], mutually intelligible with Hindi, is not present as a sample lect.) They include not only sister lects belonging to the (Indo-Aryan branch of) the Indo-European family, but also the other lects spoken across South Asia, namely Dravidian, Austroasiatic, Sino-Tibetan, and the lect isolates Nihali [niha1238] (Nagaraja 2014) and Burushaski [buru1296] (Yoshioka 2012). This clearly

demonstrates the high degree of convergence between Hindi and other lesser-spoken lects of India, although the directionality of convergence is less clear.

4 Clustering

Based on the phonological distances, we cluster similar lects into a few groups to investigate areal patterns. We use two clustering algorithms: *k*-means clustering and divisive analysis clustering.

k-means clustering (Lloyd 1982) is a statistical method to divide observations into *k* number of clusters such that the variance within each cluster is minimized. Based on the distances of each lect from other lects, we can group the lects into any number of clusters to see whether the lects are clustered geographically on a map of Eurasia, thereby forming phonological areas. Since fewer clusters are more reliable than larger number of clusters, here we only show the first three numbers of clustering: two (Figure 1), three (Figure 2), and four (Figure 3).

From the visualized *k*-means clustering, we observe that phonological clusters also tend to form geographical clusters, confirming the prediction of the areality of

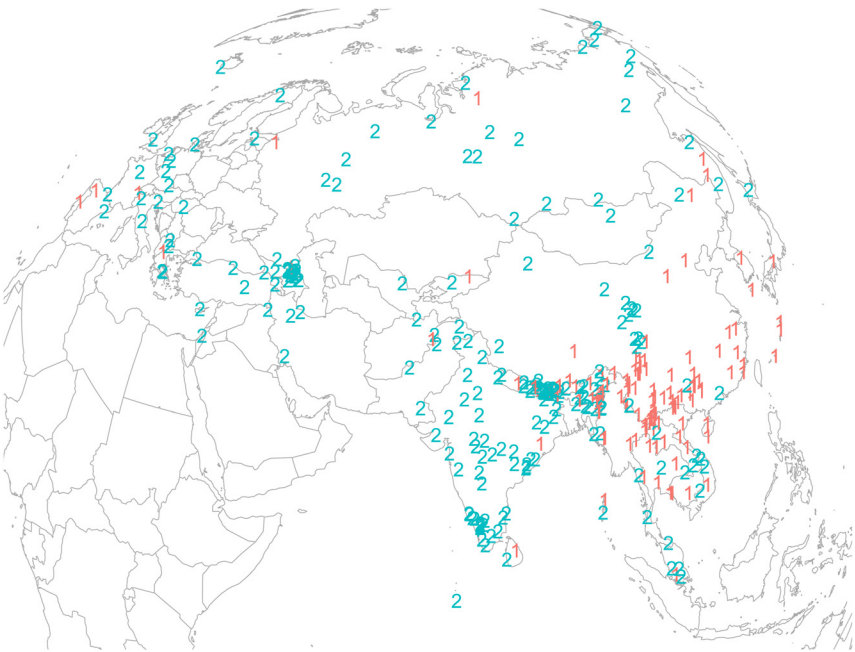


Figure 1: Two phonological clusters of Eurasia.

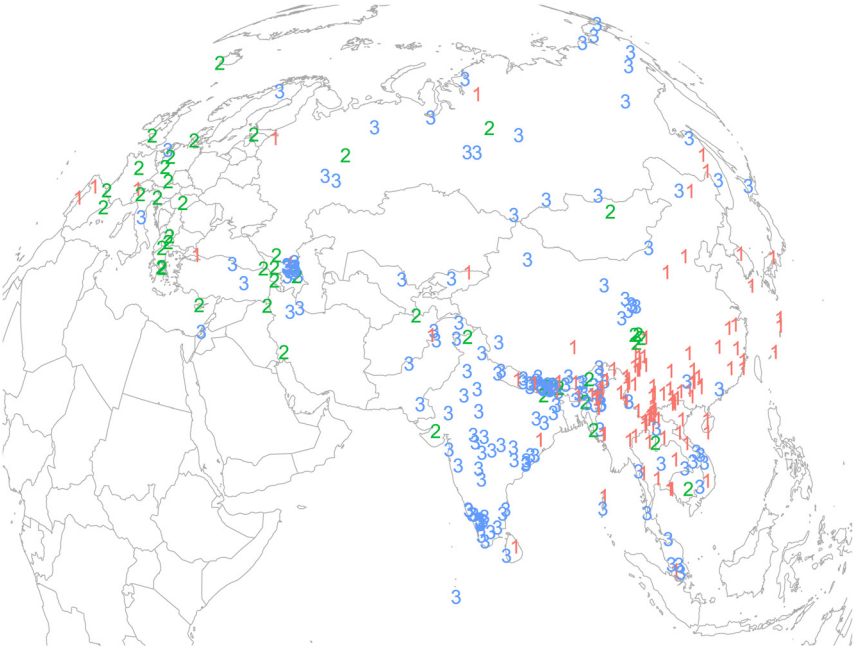


Figure 2: Three phonological clusters of Eurasia.

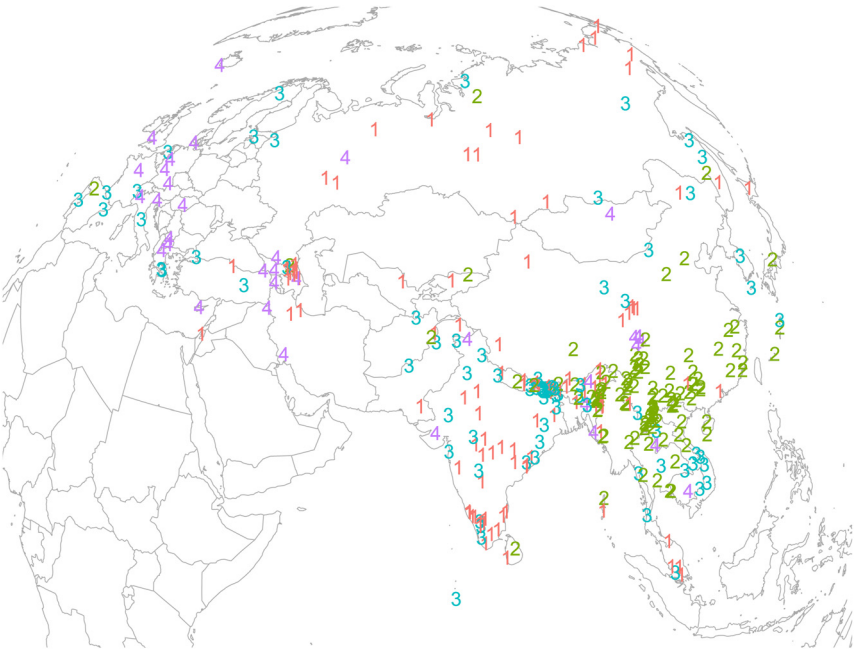


Figure 3: Four phonological clusters of Eurasia.

phonological convergence among the lects of Eurasia. The clusters do show some degree of areal inconsistency, however. It appears to be at four clusters (Figure 3) that areal inconsistency becomes considerably high, suggesting that the optimal number of clusters is three (Figure 2). The ternary clustering horizontally divides Eurasia into three clusters: eastern (i.e. East/Southeast Asia), central (i.e. South/North/Central/West Asia), and western (i.e. Europe). This ternary division may be a useful way to frame the areal landscape of the Eurasian macroarea.

Unlike *k*-means clustering, divisive analysis clustering divides the samples into groups until every sample is divided from each other. The Appendix shows a dendrogram of the hierarchically clustered Eurasian sample lects using the this algorithm. This clustering also reflects a ternary division of east/central/west: The first branching divides eastern Eurasia from the rest of Eurasia, while the second branching divides the rest of Eurasia into Europe and central Eurasia. (The second branching of eastern Eurasia does not appear to be areally consistent.)

5 Multidimensional scaling

We have also performed multidimensional scaling (k = 3) to visualize the distance between Eurasian lects. It is available at `ianjoo.github.io/PhonoMDS.html` as an interactive map.

It should be noted that the more syllabically complex a lect is, the longer the overall distance between that lect and others tend to be. In other words, due to the nature of the distance calculation method, lects with simple syllabic structures tend to have a shorter average distance from all other lects, whereas lects with complex syllabic structures tend to have longer average distance from all other lects. This is because the more possible onset/nucleus/coda forms a lect has, the higher the probability that those forms are distinct from the onset/nucleus/coda forms of other lects. This explains why the Mainland Southeast Asian lects, known for their simple syllabic forms, are tightly clustered, whereas the Eurasian lects, whose syllabic forms tend to be complex, are scantly clustered in the multidimensional scaling.

From the multidimensional scaling we can see that Georgian [nucl1302] (Butskhrikidze 2002) and Aromanian [arom1237] (Caragiu-Marioţeanu 1968) are clear outliers. Georgian is well known for allowing a vast amount of onset clusters, including typologically rare ones, such as triconsonantal plosive sequences or plosive-fricative-plosive sequences. While Laz [lazz1240] (Lacroix 2009), the other Kartvelian lect in the database, also has a long list of onset clusters, Georgian has a lot more. In addition, because the triconsonantal and quadriconsonantal sequences were not exhaustively listed in Butskhrikidze (2002), they were transcribed in underspecified segment symbols (e.g. <FPR> for fricative-plosive-sonorant

sequences), which in our analysis were converted into all logically possible sequences based on the Georgian phonemic inventory (e.g. <FPR> converted into all logically possible sequences of Georgian fricative, plosive, and sonorant phonemes). This process makes the list of Georgian complex onsets even more diverse in our analysis, although to what degree this converting algorithm contributed to increasing the diversity is unclear. The great diversity of its complex onsets (which was further increased by our algorithm at least to some degree) may have placed Georgian as an outlier from other lects of Eurasia.

As for Aromanian, its outlyingness may be partially due to the consulted source (Caragiu-Marioţeanu 1968) analyzing it as completely codaless, analyzing all word-medial clusters as onset clusters instead of dividing them into codas and onsets, including clusters such as /mbl/ or /ntr/. Such onset clusters are typologically rare, as they violate the sonority sequencing principle (Clements 1990). Moreover, being codaless is by itself somewhat rare, only attested among certain Southeast Asian lects elsewhere in Eurasia, such as some Hmong-Mien lects (Hmong Njua [hmon1264], Harriehausen 1990; Western Xiangxi Miao [west2430], Sposato 2015). While we have mostly followed the consulted literature's analysis and have coded Aromanian as codaless, had we consulted a source that analyzes Aromanian as having codas (e.g., Nisioi 2014), the phonological distances measured between Aromanian and other lects may have been considerably shorter.

6 Machine-learning prediction of linguistic areas

In this section, we use the phonological distances between Eurasian lects to test the previous hypotheses on the linguistic areas of Eurasia. A linguistic area (also known as a SPRACHBUND) is a geographical area home to multiple lects that share a number of linguistic features due to historical contact and not genealogical relationship. In other words, it is a geographical group of linguistic convergence.

Six major linguistic areas of Eurasia often discussed in the literature are:

- Northeast Asia (Whitman (Hölzl 2018; Whitman 2016))
- Qinghai-Gansu (Dwyer 2013; Janhunen 2006; Xu 2017)
- Mainland Southeast Asia (Comrie 2007; Enfield 2018; Sidwell and Jenny 2021; Vittrant and Watkins 2019)
- South Asia (Abbi 2018; Emeneau 1956, 1969; Masica 2005)
- Europe (Haspelmath 1998, 2001)
- Caucasus (controversial – proponents include Chirikba 2008; Daniel and Lander 2011; opponents include Catford 1977; Tuite 1999)

While there are certainly many more linguistic areas in Eurasia previously argued for, most famously the Balkan sprachbund (Mišeska Tomić 2006), the six major areas above are, according to the areal literature so far, the highest layers of linguistic areas and not those that form a subset of these six (such as the Balkans, which is part of Europe).

To test the validity of these linguistic area hypotheses (at least in the phonological domain), we will examine how well machine learning predicts the area of a lect given its phonological distance from other lects. For example, based on how similar German is to other Eurasian lects, can we predict that it is spoken in Europe? If machine learning, when given the phonotactic details of half of the sample lects hypothetically belonging to a linguistic area, can predict well which sample lects are the other half of the sample lects belonging to the same area, we will be able to conclude that the information provided by Phonotacticon is sufficient to make generalizations on areal similarities, validating the claim that the phonotactic information coded in Phonotacticon can demonstrate linguistic areahood (in the phonological domain).

We first divide the Eurasian lects into six different regions solely based on their geographical coordinates: Northeast Asia, Mainland Southeast Asia, Qinghai-Gansu, South Asia, West Asia, and Europe. We exclude the Caucasus due to its controversial status as a linguistic area and the fact that the clustering and multidimensional scaling analyses (Sections 4 and 5) do not seem to suggest that Caucasian lects tend to cluster. While West Asia has not been frequently discussed as a potential linguistic area, it is nevertheless posited as a region so that every Eurasian lect will belong to one region. Figure 4 visualizes the lects in the predefined seven regions.

The goal is to train a model based on phonological distance to see how well it predicts in which of these six areas a given lect is spoken. We train the naive Bayes classifier based on half of the lects and their distance from other lects. First, we divide the sample in half by area. (The proportion of the areas is thus equal in the halved sample.) Then we train the classifier in the first half and test it on the other half. We also perform the opposite by training the second half and testing it on the first half. Finally, we join the two halves.

The results show that the accuracy of the predictions (0.61) is significantly higher than the No Information Rate (0.42, $p < 0.001$), suggesting that the naive Bayes classifier can predict whether a given lect belongs to a pre-defined linguistic area with a significant level of accuracy. The kappa value (0.47) also shows that the model successfully predicts the areas to a moderate degree.

The F1 values of individual classes (Table 9) show that the model predicts some areas better than others. Mainland Southeast Asia and South Asia are predicted well and Qinghai-Gansu only poorly so. This may be partly because of the uneven sample size across different regions, Qinghai-Gansu being the smallest with only nine

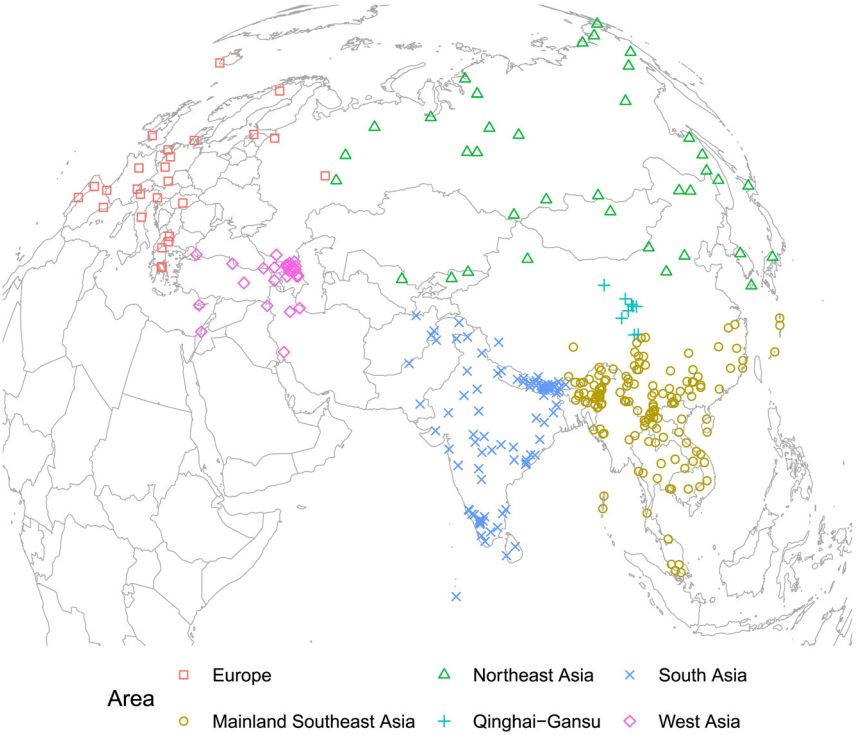


Figure 4: Sample lects divided by geographically pre-defined regions of Eurasia.

sample lects. West Asia, although not previously discussed as a linguistic area, is predicted quite reliably.

Figure 5 is the visualization of the lects by their predicted areas. Overall, the naive Bayes classifier predicts well the six linguistic areas, not including Qinghai-Gansu, which may be due to the small sample size.

Table 9: F1 values of individual classes.

Class	F1
Europe	0.45
Mainland Southeast Asia	0.77
Northeast Asia	0.47
Qinghai-Gansu	0.27
South Asia	0.60
West Asia	0.43

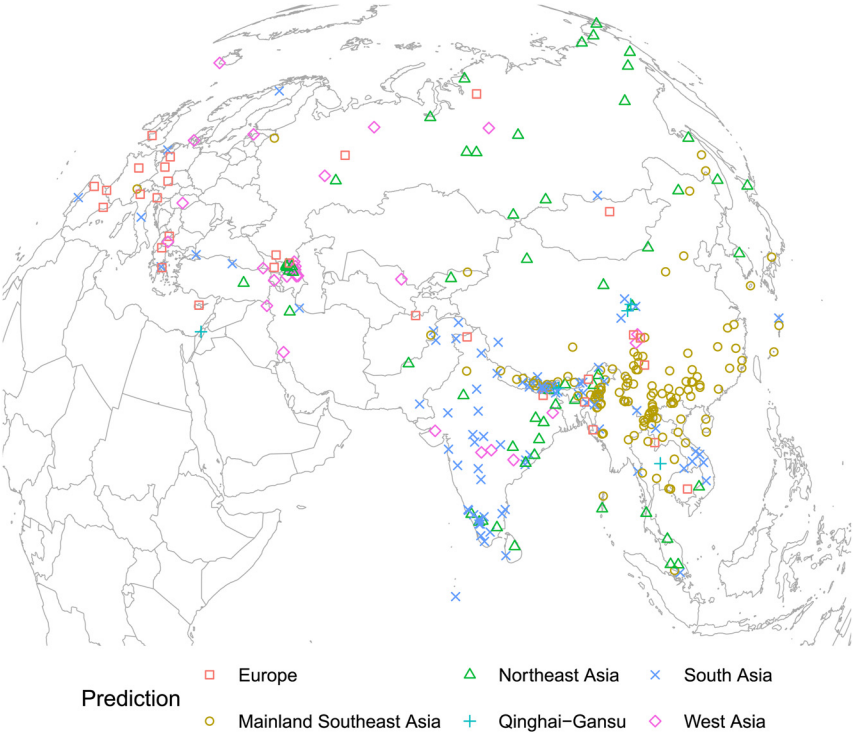


Figure 5: Map of sample lects and the areas they were predicted to belong to.

7 Testing the correlation between geographical distance and phonological distance

We will also test the hypothesis that geographical distance is correlated with phonological distance; in other words, the closer two lects are geographically, the closer they tend to be phonologically.

In order to test this hypothesis, we use Burushaski [buru1296] as a point of reference. Burushaski is a lect isolate spoken by the Burusho people in the northernmost valleys of Pakistan (Yoshioka 2012). As it is genealogically an isolate and geographically located in the central part of Eurasia, it can be an optimal scale to measure the correlation between geographical and phonological distance. The refined hypothesis is thus: the closer a Eurasian lect to Burushaski geographically, the closer it will be phonologically.

As shown in Section 4, geographically close lects also tend to show similar phonological patterns. Thus, spatial autocorrelation must be tested for the lects' distances to Burushaski. First, we define the geographical neighbors of each lect. The neighbors of each lect are defined as the lects whose geographical coordinates are within the maximal distance of any lect to its nearest lect. In other words, we defined neighborhood such that every lect will have at least one geographical neighbor.

We then generate a spatial weight matrix based on binary weight. The value of 1 is assigned to each neighboring-lect pair, while the value of 0 is assigned to other pairs. The rows are then standardized so that all the neighbor weights of each lect add up to 1. Moran's I test based on this spatial weight matrix validates the spatial autocorrelation ($p < 0.001$). Finally, using this spatial weight matrix, we perform a spatial regression analysis based on the spatial lag model. The results show that geographical distance to Burushaski is not significantly correlated with the phonological distance to it ($p > 0.05$). The phonological distance of each lect to Burushaski, rounded to the closest integer, is visualized in Figure 6. Based on this observation, we fail to draw the conclusion that phonological distance is correlated with geographical distance in Eurasia.

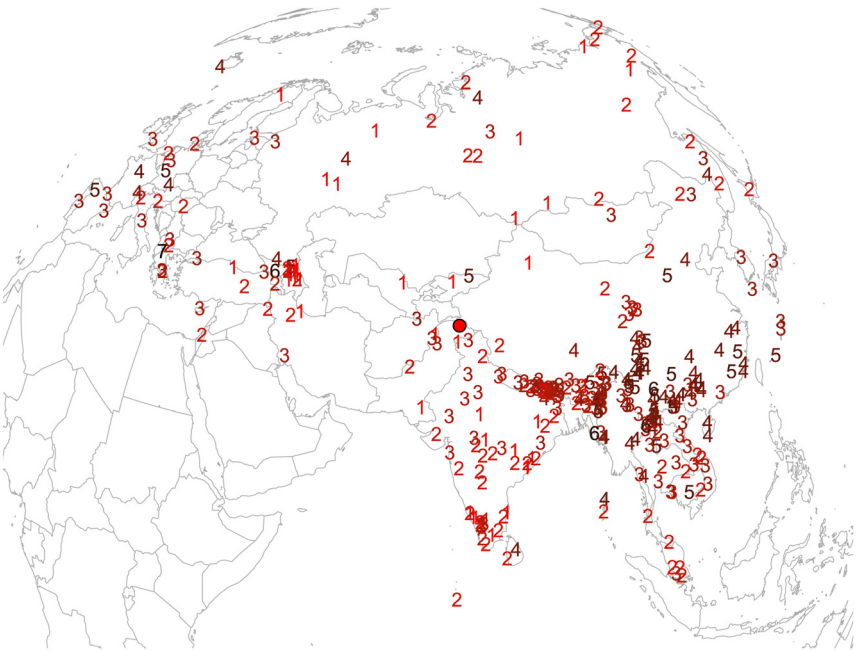


Figure 6: Each Eurasian lect's phonological distance from Burushaski, rounded to the closest integer.

8 Comparison with morphosyntactic distance

Convergence may be domain-specific; that is, convergence in one domain does not entail convergence in other domains. Meakins and Pensalfini (2021) show that two Australian lects, Jingulu [djin1251] and Mudburra [mudb1240], share a great deal of mutually borrowed vocabulary but retain each of their distinct grammar. François (2011) demonstrates how northern Vanuatu lects (all belonging to the Oceanic branch of the Austronesian family) have phonologically and lexically diverged but show a great degree of syntactic isomorphism. Donohue (2013: 223) takes Basque and Dravidian lects as examples where the dominant Indo-European lects have affected their phonology and Khoi-San as an example of lects having received morphosyntactic influence from the Niger-Congo superstratum. Thus, the phonological convergence patterns we have seen so far do not necessarily imply that similar morphosyntactic convergence has emerged.

However, given that phonological convergence is mostly motivated by contact and that lects in contact are likely to converge in both phonological and morphosyntactic domains, it is plausible that similar morphosyntactic convergence patterns have occurred throughout Eurasia. The two domains of areality could be parallel in some regions of Eurasia and differ from each other in other regions, which would both be equally interesting.

In order to compare the phonological areal patterns to morphosyntactic areal patterns, we measure the morphosyntactic distance between Eurasian lects using Grambank 1.0 (Skirgård et al. 2023). Grambank 1.0 is a database consisting of 2,457 lects and their values of 195 morphosyntactic features, of which 189 are binary parameters (e.g. *Is there a gender distinction in independent 3rd person pronouns?*). As it does not contain information on phonological features, Phonotacticon can serve as a good complement to Grambank. The two databases used in comparison can cross-validate areal patterns in Eurasia or discover domain-specific patterns.

Based on the 189 binary features of Grambank, we calculated the Manhattan distance between each pair of lects, whose vectors consist of 1 (positive), −1 (negative), or 0 (unknown) values of each feature. (Ca. 13 % of all feature values are marked as unknown.) Then, based on *k*-means clustering, we clustered the Eurasian lects into two, three, and four clusters, visualized in Figures 7–9.

The clusterings based on Grambank are much more neatly defined compared to the those based on Phonotacticon. One reason for this difference could be that, as we will discuss in Section 9, morphosyntactic features may be genealogically more conservative than phonological features. The genealogical conservatism may have led the lects in the same family to be clustered together (which also tend to be distributed in the same geographical region), resulting in geographically more consistent clusters.

It may leave the impression to the reader that the morphosyntactic clusters are somewhat similar to the phonological clusters visualized in Figures 1–3. What is

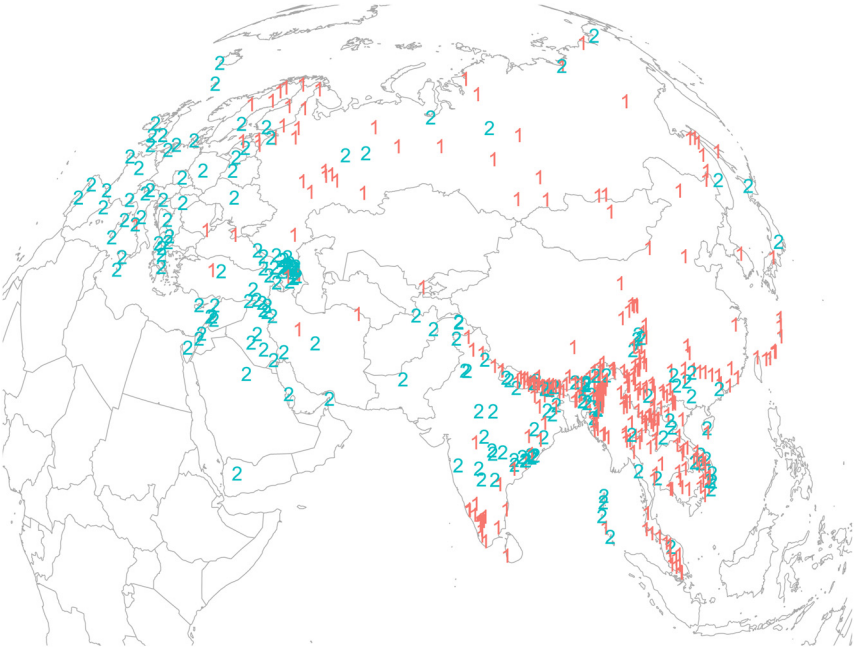


Figure 7: Two morphosyntactic clusters of Eurasia.

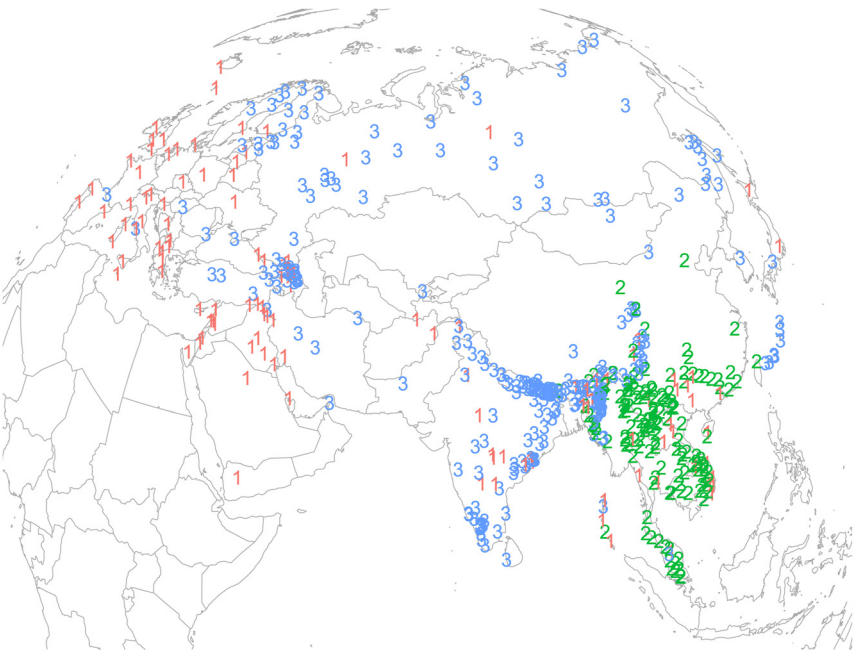


Figure 8: Three morphosyntactic clusters of Eurasia.

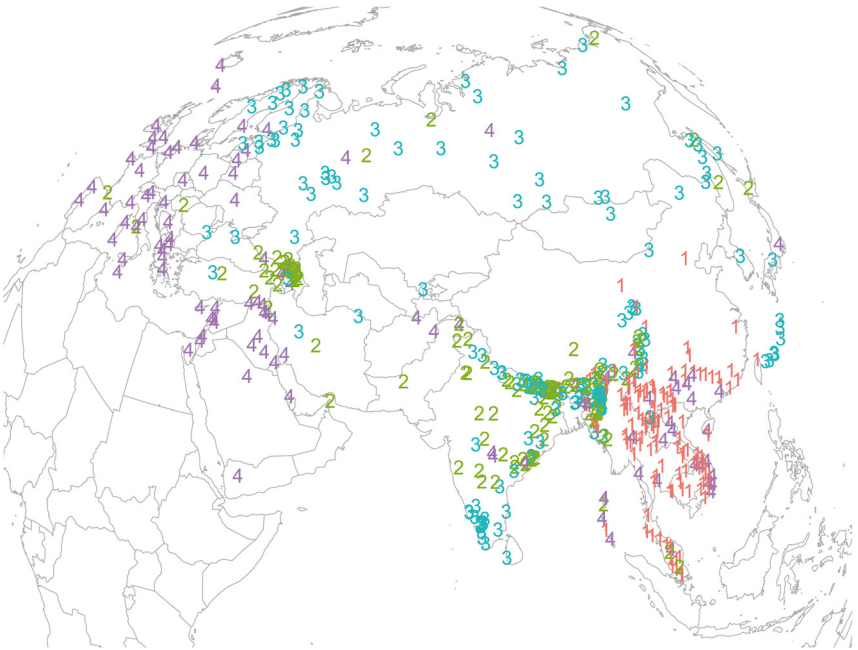


Figure 9: Four morphosyntactic clusters of Eurasia.

striking is that the morphosyntactic grouping also shows a horizontal ternary division (Figure 8), even when the number of clusters is increased to four (Figure 9). Even though the boundaries of the three clusters do not match exactly between phonological and morphosyntactic groupings, the preference for ternary division shown in both sets of *k*-means clusterings suggests that Eurasian lects may be best divided into three groups: eastern, central, and western.

In order to test the similarity between the phonological distances and morphosyntactic distances, we ran a Mantel test based on Pearson's correlation coefficient to test the correlation between the two distance matrices. When limiting the sample lects to the intersection between Phonotacticon and Grambank, or 222 lects, the Mantel statistic R is 0.215 from −1 to 1 scale ($p < 0.001$), −1 representing a perfectly negative correlation and 1 a perfectly positive correlation. This shows a moderate level of correlation between phonological distance and morphosyntactic distance between Eurasian lects.

Thus, while phonological similarities and morphosyntactic similarities overlap to some degree, they do not always match, which aligns with previous findings that phonological convergence between two lects may not imply their convergence in other domains, or vice versa. The difference between phonological and morphosyntactic clusterings reflects the domain-specific nature of linguistic convergence, although the preference for ternary division shared by both clusterings merits further investigation.

9 Comparison with genealogy

Lastly, we will test whether the phonological distances measured via Phonotacticon are correlated to the genealogical distances between Eurasian lects. Linguistic similarity arises not only due to historical contact but also genealogical relatedness. Within a family, the more genealogical layers shared by two lects, the shorter we can expect the phonological distances between them to be. For example, French [stan1290] and Italian [ital1282] not only belong to the same Indo-European family, but also share seven layers within Indo-European (Classical Indo-European > Italic > Latino-Faliscan > Latinic > Imperial Latin > Romance > Italo-Western Romance), based on Glottolog 4.4. Due to such close genealogy, we can predict that the phonological distance between French and Italian will be significantly shorter than that between French and Welsh [wels1247], which only shares one layer (Classical Indo-European) whence they diverged at a much earlier time. More generally, we can predict that within a family, the phonological distance between two lects will decrease as the number of shared layers increases.

In order to verify this prediction, we tested Pearson's correlation coefficient between the number of layers shared by two lects and their phonological distance per family. The null hypothesis is that the number of shared layers and the phonological distance are not correlated to each other at all, i.e. Pearson's correlation coefficient will not be significantly different from 0 (absolute absence of correlation). The alternative hypothesis is that the number of shared layers and the phonological distance are correlated, making Pearson's correlation coefficient significantly different from 0, given that the sample size (the number of lect pairs) is large enough. To correct for multiple comparisons, the *p*-values of Pearson's correlation coefficients are adjusted as false discovery rates (Benjamini and Hochberg 1995), the threshold for significance being 0.1. Families that have no internal layers and/or less than three sample lects (such as Koreanic, whose only two members are Korean [kore1280] and Jejueo [jeju1234] without any internal layer) are excluded, as they lack the minimal number of shared layers and lect pairs needed to test the correlation.

The results shown in Table 10 fail the prediction and uphold the null hypothesis. Among the tested Eurasian families, no family shows a statistically significant correlation (FDR = 0.1) between phonological distance and genealogical relatedness. This is true not only for families with a small number of lect pairs, such as Japonic or Hmong-Mien, but also for those with a sizeable sample, such as Indo-European and Sino-Tibetan. This suggests that the phonological distances between Eurasian lects measured via Phonotacticon must have not been shaped strongly by genealogical heritage.

Table 10: Pearson's correlation coefficient (r) and the false discovery rate (FDR) between the phonological distance and the number of shared genealogical layers between two lects of the same family.

Family	Number of lect pairs	r	FDR
Nakh-Daghestanian	78	−0.19	1.00
Turkic	66	−0.18	1.00
Tai-Kadai	153	−0.09	1.00
Dravidian	210	−0.07	1.00
Indo-European	1,953	−0.03	1.00
Sino-Tibetan	7,503	−0.01	1.00
Austroasiatic	351	−0.00	1.00
Mongolic-Khitan	28	0.01	1.00
Uralic	36	0.12	1.00
Japonic	6	0.13	1.00
Tungusic	15	0.34	1.00
Austronesian	21	0.39	1.00
Hmong-Mien	10	0.42	1.00

Is this correlation also absent in morphosyntax? In order to make a comparison between the two domains of phonology and morphosyntax, we also tested the correlation between the number of shared layers and morphosyntactic distance measured using Grambank (Section 8). Table 11 shows the correlation between the number of shared layers and morphosyntactic distances between the same sample

Table 11: Pearson's correlation coefficient (r) and the false discovery rate (FDR) between the morphosyntactic distance and the number of shared genealogical layers between two lects of the same family.

Family	Number of lect pairs	r	FDR
Japonic	15	−0.75	0.01
Uralic	300	−0.57	0.00
Dravidian	435	−0.51	0.00
Indo-European	2,016	−0.46	0.00
Turkic	91	−0.37	0.00
Tungusic	36	−0.28	0.57
Sino-Tibetan	17,020	−0.27	0.00
Mongolic-Khitan	28	−0.21	0.78
Austroasiatic	3,240	−0.19	0.00
Tai-Kadai	120	−0.14	0.63
Nakh-Daghestanian	210	−0.13	0.42
Afro-Asiatic	231	−0.07	0.78
Hmong-Mien	45	0.05	0.78
Austronesian	120	0.12	0.78

lects. The results are strikingly different from those from Phonotacticon (Table 10): when corrected for multiple comparisons at false discovery rate of 0.1, seven out of fourteen families show a negative correlation between the number of shared layers and morphosyntactic distance. This suggests that morphosyntactic distances may be more heavily related to genealogical heritage compared to phonological distances.

10 Conclusions

In this paper, we have measured the phonological distance between Eurasian lects using Phonotacticon 1.0. Importantly, the distance measuring is based on the entirety of the phonotactic data available in Phonotacticon, except for the tonal qualities (as only the number of tones was used to calculate the distance between tonal inventories). Here we summarize our findings:

- Clustering the lects together based on phonological distance shows that phonologically close lects also tend to be geographically close and suggests a ternary division of eastern/central/western Eurasia.
- On the other hand, comparison between the geographical and the phonological distances to Burushaski shows that geographical and phonological distances do not significantly correlate with each other.
- Machine learning based on the phonological distances can predict the linguistic area of each lect to a moderate degree.
- Comparing phonological clustering to morphosyntactic clustering based on Grambank shows that the convergence patterns in the two domains have only a moderate degree of similarity, suggesting that linguistic convergence may show different areal patterns in different domains, although both clusterings share a preference for a ternary division.
- The absence of correlation between phonological distance and the number of shared genealogical layers as well as the presence of such correlation in morphosyntactic distance suggests that morphosyntax may more strongly preserve its genealogical heritage compared to phonology.

The methodology employed to measure interstructural phonological distance is of course limited in many aspects and just one of many possible methodologies. There are two levels of limitations: the limitation of the data and the limitation of the analysis. First, as mentioned in Section 3.1, while Phonotacticon contains a large and important part of each lect's phonology, it certainly does not capture the entirety of it, such as morphophonemic processes or phonotactic restrictions beyond syllabic boundaries. The phonological distances could be measured in finer resolution with

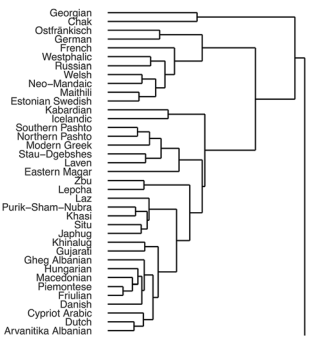
data containing a broader range of phonological information. Second, as mentioned in Section 3.6, some parts of the distance analysis is rather simplistic, such as assigning equal weight to all four normalized distances (onset/nucleus/coda/tone).

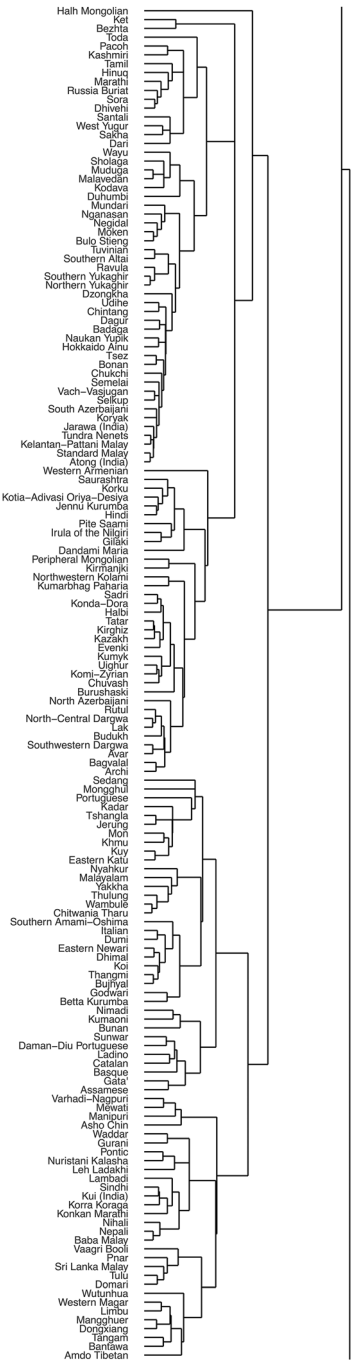
While there is much room for improvement, the present study is one of the first steps towards quantifying interstructural phonological distance based on large-scale cross-linguistic data and thereby discovering areal patterns. Further exploration of this task is needed, not only using Phonotacticon but also other databases that may be created in the future.

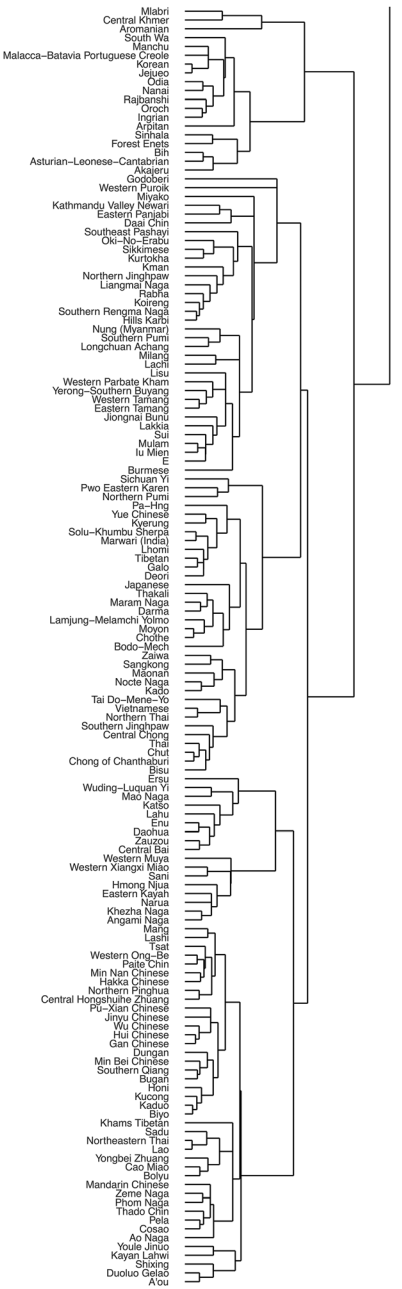
Acknowledgments: We would like to thank Chor Shing David Li (the Hong Kong Polytechnic University) for his supervision of the first author of this paper as co-supervisor. Our gratitude also goes to Harald Hammarström (Uppsala University), who was the host supervisor of the first author during his three-month visit to his institution. Lastly, we are grateful for the anonymous reviewers for their constructive feedback.
Author contributions: I. J. collected the data, reviewed the previous literature, and conducted the visualizations and statistical analyses. Y.-Y. H. provided constant feedback as I. J.'s doctoral supervisor. All authors reviewed the manuscript.
Research funding: This work was supported by JSPS KAKENHI Grant Number 24K23937.

Appendix

The following dendrogram shows a divisive analysis clustering of the Eurasian sample lects based on their phonological distances, divided into western, central, and eastern Eurasia by page.




Halh Mongolian
Ket
Bezhta
Toda
Pacoh
Kashmiri
Tamil
Hinuq
Marathi
Russia Buriat
Sora
Dhivehi
Santali
West Yugur
Sakha
Dari
Wayu
Sholaga
Muduga
Malavedan
Kodava
Duhumbi
Mundari
Nganasan
Negidal
Móken
Bulg Stieng
Tuvinian
Southern Altai
Ravula
Southern Yukaghir
Northern Yukaghir
Dzongkha
Udihe
Chintang
Dagur
Badaga
Naukan Yupik
Hokkaido Ainu
Tsez
Bonan
Chukchi
Semelai
Vach–Vasiugan
Selkup
South Azerbaijani
Koryak
Jarawa (India)
Tundra Nenets
Kelantan–Pattani Malay
Standard Malay
Atong (India)
Western Armenian
Saurashtra
Korku
Kotia–Adivasi Oriya–Desiya
Jennu Kurumba
Hindi
Pite Saami
Irula of the Nilgiri
Giliaki
Dandami Maria
Peripheral Mongolian
Kirmanjki
Northwestern Kolami
Kumarbhag Paharia
Sadri
Konda–Dora
Halbi
Tatar
Kirghiz
Kazakh
Evenki
Kumyk
Uighur
Komi–Zyrian
Chuvash
Burushaski
North Azerbaijani
Rutul
North–Central Dargwa
Lak
Budukh
Southwestern Dargwa
Avar
Bagvalal
Archi
Sedang
Monghul
Portuguese
Kadar
Tshangla
Jerung
Mon
Khmu
Kuy
Eastern Katu
Nyahkur
Malayalam
Yakkha
Thulung
Wambule
Chitwania Tharu
Southern Amami–Oshima
Italian
Dumi
Eastern Newari
Dhimal
Koi
Thangmi
Bujhyal
Godwari
Betta Kurumba
Nimadi
Kumaoni
Bunan
Sunwar
Daman–Diu Portuguese
Ladino
Catalan
Basque
Gata'
Assamese
Varhadi–Nagpuri
Mewati
Manipuri
Asho Chin
Waddar
Gurani
Pontic
Nuristani Kalasha
Leh Ladakhi
Lambadi
Sindhi
Kui (India)
Korra Koraga
Konkan Marathi
Nihali
Nepali
Baba Malay
Vaagri Booli
Pnar
Sri Lanka Malay
Tulu
Domari
Wutunhua
Western Magar
Limbu
Mangghuer
Dongxiang
Tangam
Bantawa
Amdo Tibetan


Mlabri
Central Khmer
Aromanian
South Wa
Manchu
Malacca–Batavia Portuguese Creole
Korean
Jejueo
Odia
Nanai
Rajbanshi
Oroch
Ingrian
Arpitan
Sinhala
Forest Enets
Bih
Asturian–Leonese–Cantabrian
Akajeru
Godoberi
Western Puroik
Miyako
Kathmandu Valley Newari
Eastern Panjabi
Daai Chin
Southeast Pashayi
Oki–No–Erabu
Sikkimese
Kurtokha
Kman
Northern Jinghpaw
Liangmai Naga
Rabha
Koireng
Southern Rengma Naga
Hills Karbi
Nung (Myanmar)
Southern Pumi
Longchuan Achang
Milang
Lachi
Lisu
Western Parbate Kham
Yerong–Southern Buyang
Western Tamang
Eastern Tamang
Jiongnai Bunu
Lakkia
Sui
Mulam
Iu Mien
E
Burmese
Sichuan Yi
Pwo Eastern Karen
Northern Pumi
Pa–Hng
Yue Chinese
Kyerung
Solu–Khumbu Sherpa
Marwari (India)
Lhomi
Tibetan
Galo
Deori
Japanese
Thakali
Maram Naga
Darma
Lamjung–Melamchi Yolmo
Moyon
Cholhe
Bodo–Mech
Zaiwa
Sangkong
Maonan
Nocte Naga
Kado
Tai Do–Mene–Yo
Vietnamese
Northern Thai
Southern Jinghpaw
Central Chong
Thai
Chut
Chong of Chanthaburi
Bisu
Ersu
Wuding–Luquan Yi
Mao Naga
Kaiso
Lahu
Enu
Daohua
Zauzou
Central Bai
Western Muya
Western Xiangxi Miao
Sani
Hmong Njua
Eastern Kayah
Narua
Khezha Naga
Angami Naga
Maring
Lashi
Tsat
Western Ong–Be
Paite Chin
Min Nan Chinese
Hakka Chinese
Northern Pinghua
Central Hongshuihe Zhuang
Pu–Xian Chinese
Jinyu Chinese
Wu Chinese
Hui Chinese
Gan Chinese
Dungan
Min Bei Chinese
Southern Qiang
Bugan
Honi
Kucong
Kaduo
Biyo
Khams Tibetan
Sadu
Northeastern Thai
Lao
Yongbei Zhuang
Cao Miao
Bolyu
Mandarin Chinese
Zeme Naga
Phom Naga
Thado Chin
Pela
Cosao
Ao Naga
Youle Jinuo
Kayan Lahwi
Shixing
Duoluo Gelao
A'ou

References

Abbi, Anvita. 2018. Echo formations and expressives in South Asian languages. In Aina Urdze (ed.), *Non-prototypical reduplication*, 1–34. Berlin: De Gruyter.

Afendras, Evangelos A. 1970. Can one measure a sprachbund? A calculus of phonemic distribution for language contact. *Folia Linguistica* 4(1–2). https://doi.org/10.1515/flin.1970.4.1-2.93.

Avram, Andrei. 1964. Sur la typologie phonologique quantitative [On the quantitative phonological typology]. *Revue Roumaine de Linguistique* IX. 131–134.

Bauer, Robert S. & Paul K. Benedict. 1997. *Modern Cantonese phonology*. Berlin & New York: Mouton de Gruyter.

Beebe, Leslie M. 1975. Occupational prestige and consonant cluster simplification in Bangkok Thai. *International Journal of the Sociology of Language* 1975(5). https://doi.org/10.1515/ijsl.1975.5.43.

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)* 57(1). 289–300.

Bhatia, Tej K. 1993. *Punjabi: A cognitive-descriptive grammar*. London: Routledge.

Bičan, Aleš. 2011. *Phonotactics of Czech*. Brno: Masaryk University Dissertation.

Booij, Geert. 1999. *The phonology of Dutch*. Oxford: Oxford University Press.

Butskhrikidze, Marika. 2002. *The consonant phonotactics of Georgian*. Leiden: Leiden University Dissertation.

Caragiu-Marioțeanu, Matilda. 1968. *Fono-morfologie aromână: studiu de dialectologie structurală [Aromanian phono-morphology: Structural dialectology study]*. Bucharest: Editura Academiei Republicii Socialiste România.

Catford, J. C. 1977. Mountain of tongues: The languages of the Caucasus. *Annual Review of Anthropology* 6(1). 283–314.

Chirikba, Viacheslav A. 2008. The problem of the Caucasian Sprachbund. In Pieter Muysken (ed.), *From linguistic areas to areal linguistics*, 25–93. Amsterdam: John Benjamins.

Clements, George. 1990. The role of the sonority cycle in core syllabification. In John Kingston & Mary Beckman (eds.), *Papers in laboratory phonology*, vol. I, 283–333. Cambridge: Cambridge University Press.

Comrie, Bernard. 2007. Areal typology of Mainland Southeast Asia: What we learn from the WALS maps. *MANUSYA: Journal of Humanities* 10(3). 18–47.

Daniel, Michael & Yury Lander. 2011. The Caucasian languages. In *The languages and linguistics of Europe: A comprehensive guide*, 125–158. Berlin: De Gruyter Mouton.

Do, Youngah & Ryan Ka Yau Lai. 2021. Accounting for lexical tones when modeling phonological distance. *Language* 97(1). e39–e67.

Donohue, Mark. 2013. Who inherits what, when? Toward a theory of contact, substrates, and superimposition zones. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency: In honor of Johanna Nichols*, 219–240. Amsterdam: John Benjamins.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS online (v2020.4)*. Zenodo.

Dwyer, Arienne. 2013. Tibetan as a dominant Sprachbund language: Its interactions with neighboring languages. In *The third international conference on the Tibetan language*, 258–280. New York: Trace Foundation.

Eden, S. Elizabeth. 2018. *Measuring phonological distance between languages*. London: University College London Dissertation.

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

Emeneau, Murray B. 1969. Onomatopoetics in the Indian linguistic area. *Language* 45(2). 274–299.

Enfield, Nick James. 2018. *Mainland Southeast Asian languages: A concise typological introduction*. Cambridge: Cambridge University Press.

François, Alexandre. 2011. Social ecology and language history in the northern Vanuatu linkage: A tale of divergence and convergence. *Journal of Historical Linguistics* 1(2). 175–246.

Gut, Ulrike. 2009. *Introduction to English phonetics and phonology*, vol. I. Frankfurt am Main: Peter Lang.

Harnud, Huhe & Xuewen Zhou. 2021. On the relation between the similarity of the acoustic distribution patterns of vowels and the language closeness. *International Journal of Anthropology and Ethnology* 5(1). 1–13.

Harriehausen, Bettina. 1990. *Hmong Njua: Syntaktische Analyse einer gesprochenen Sprache mithilfe datenverarbeitungstechnischer Mittel und sprachvergleichende Beschreibung des südostasiatischen Sprachraumes [Hmong Njua: Syntactic analysis of a spoken language using data processing technology and a comparative description of the Southeast Asian linguistic area]*. Tübingen: Max Niemeyer.

Haspelmath, Martin. 1998. How young is Standard Average European? *Language Sciences* 20(3). 271–287.

Haspelmath, Martin. 2001. The European linguistic area: Standard Average European. In Martin Haspelmath (ed.), *Language typology and language universals/Sprachtypologie und sprachliche Universalien/La typologie des langues et les universaux linguistiques*, vol. II, 1492–1510. Berlin: De Gruyter Mouton.

Hölzl, Andreas. 2018. *A typology of questions in Northeast Asia and beyond: An ecological perspective*. Berlin: Language Science Press.

Iwasaki, Shoichi & Preeya Ingkaphirom. 2005. *A reference grammar of Thai*. New York: Cambridge University Press.

Janhunen, Juha. 2006. Sinitic and non-Sinitic phonology in the languages of Amdo Qinghai. In Christoph Anderl & Eifring Halvor (eds.), *Studies in Chinese language and culture: Festschrift in honour of Christoph Harbsmeier on the occasion of his 60th birthday*, 261–268. Oslo: Hermes Academic Publishing.

Jenny, Mathias & San San Hnin Tun. 2016. *Burmese: A comprehensive grammar*. London: Routledge.

Joo, Ian & Yu-Yin Hsu. 2025. Phonotacticon: A cross-linguistic phonotactic database. *Linguistic Typology* 29(2). 405–431.

Kachru, Yamuna. 2006. *Hindi*. Amsterdam: John Benjamins.

Kongsin, Phramaha Kham-Iang. 1979. *A descriptive analysis of the modern Northeastern Thai dialect (Modern Thai I'san)*. Pune: Deccan College Dissertation.

Kučera, Henry & George K. Monroe. 1968. *A comparative quantitative phonology of Russian, Czech, and German*. New York: American Elsevier.

Lacroix, René. 2009. *Description du dialecte laze d'Arhavi (caucasique du sud, Turquie): Grammaire et textes [Description of the Laz dialect of Arhavi (South Caucasian, Turkey): Grammar and texts]*. Lyon: Université Lumière Lyon 2 Dissertation.

Lloyd, Stuart. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2). 129–137.

Macklin-Cordes, Jayden L., Claire Bowern & Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38(2). 210–258.

Mao 毛, Zongwu 宗武. & Yunbing 云兵. Li 李. 2002. *Jiongnaiyu yanjiu* 炯奈语研究 *[A study of Jiongnai]*. Beijing 北京: Central Nationalities University Press 中央民族大学出版社.

Masica, Colin P. 2005. *Defining a linguistic area: South Asia*. New Delhi: Chronicle Books.

Meakins, Felicity & Rob Pensalfini. 2021. Holding the mirror up to converted languages: Two grammars, one lexicon. *International Journal of Bilingualism* 25(2). 425–457.

Mišeska Tomić, Olga. 2006. *Balkan sprachbund morpho-syntactic features*. Dordrecht: Springer.

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. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 3475–3484.

Nagaraja, K. S. 2014. *The Nihali language (grammar, texts, vocabulary)*. Mysore: Central Institute of Indian Languages.

Nikolaev, Dmitry. 2019. Areal dependency of consonant inventories. *Language Dynamics and Change* 9(1). 104–126.

Nisioi, Sergiu. 2014. On the syllabic structures of Aromanian. In *Proceedings of the 8th workshop on language technology for cultural heritage, social sciences, and humanities (LaTeCH)*, 110–118. Gothenburg: Association for Computational Linguistics.

Postovalova Постовалова, V. I. В. И. 1966. O sočetajemosti differencial'nyx priznakov soglasnyx fenomen sovremennogo russkogo jazyka О сочетаемости дифференциальных признаков согласных фонем современного русского языка [On the compatibility of the differential features of consonantal phonemes of contemporary Russian]. In E. A. Э. A. Makajev Макаев (ed.), *Problemy lingvističeskogo analiza: Fonologija, grammatika, leksikologija Проблемы лингвистического анализа: Фонология, грамматика, лексикология [Problems of linguistic analysis: Phonology, grammar, lexicology]*, 34–46. Moscow Москва: Nauka Наука.

Riad, Tomas. 2013. *The phonology of Swedish*. Oxford: Oxford University Press.

Saporta, Sol. 1955. Frequency of consonant clusters. *Language* 31(1). 25.

Sidwell, Paul & Mathias Jenny. 2021. *The languages and linguistics of Mainland Southeast Asia: A comprehensive guide*. Berlin: De Gruyter Mouton.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Heer Leonard, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175.

Sposato, Adam. 2015. *A grammar of Xong*. Amherst, NY: State University of New York at Buffalo Dissertation.

Tambovtsev, Yuri A. 2001. The phonological distances between Mongolian and Turkic languages based on typological consonantal features. *Mongolian Studies* 24. 41–84.

Tuite, Kevin. 1999. The myth of the Caucasian Sprachbund: The case of ergativity. *Lingua* 108(1). 1–29.
Vittrant, Alice & Justin Watkins (eds.). 2019. *The Mainland Southeast Asia linguistic area*. Berlin: De Gruyter Mouton.
Whitman ホイットマン, John ジョン. 2016. Tōhoku ajia gengo chiikino ichi dzukeni mukete 東北アジア言語地域の位置付けに向けて [On the Northeast Asia as a linguistic area]. 国語研プロジェクトレビュー= *NINJAL Project Review* 6. 69–82.
Xu, Dan. 2017. *The Tangwang language: An interdisciplinary case study in northwest China*. Cham: Springer.
Yoshioka, Noboru. 2012. *A reference grammar of Eastern Burushaski*. Tokyo: Tokyo University of Foreign Studies Dissertation.