Phonotacticon: a cross-linguistic phonotactic database

Ian Joo [tɕʰù.ì.an][1] and Yu-Yin Hsu [ɕy²¹.jou⁵¹.jin²¹⁴][2]

[1]Otaru University of Commerce
[2]The Hong Kong Polytechnic University

Preprint as of August 30, 2024

Abstract

Phonotacticon is a cross-linguistic database that contains syllabic phonotactic information about spoken lects (linguistic varieties), including the possible forms of the onset, nucleus, and coda of each lect, as well as the phonemic and tonemic inventories. In this paper, we present Phonotacticon 1.0, which contains the phonotactic profiles of 516 Eurasian lects retrieved from descriptive literature. The later versions of Phonotacticon will extend beyond Eurasia and will ultimately cover spoken lects in all macroareas. As an example of the research potential of this database in future studies, we have generated from Phonotacticon several descriptive visualizations, such as the distribution of the maximal onset length, to demonstrate the visually discernible areal distribution of certain phonotactic patterns.

1 Introduction

Constructing quantitative typological databases is one of the first crucial steps in enabling research on quantitative typology. In recent years, we have witnessed an increase in large-scale databases containing cross-linguistic data in different linguistic domains, such as morphosyntax (Bickel et al. 2022; Skirgård et al. 2023), lexical semantics (Rzymski et al. 2020), wordlists (List et al. 2022), and segmental phonology (Moran & McCloy 2019). Such databases are multipurpose by nature and have paved the way for diverse data-driven approaches to linguistic diversity and universals, such as the areality of sound change (Nikolaev 2019), the association between lexical form and meaning (Blasi et al. 2016), and the correlation between human physiology and sound systems (Blasi et al. 2019).

The most fruitful type of typological database that has been developed at present is the phonological database. As we will show in Section 2, most of the existing phonological databases have focused on coding the phonemic inventories of different lects. (LECT refers to any level of linguistic variety, commonly referred to as a DIALECT or a LANGUAGE). While the phonemic inventory is undoubtedly an essential part of a lect's phonology, it remains a small subset of the entire phonological profile of a lect, which includes other patterns, such as phonotactic constraints that determine how these phonemes can be distributed in relation to one another.

In order to codify a wider set of phonological patterns into a database, we have constructed the first version of PHONOTACTICON, a database that consists of the syllabic phonotactic information of spoken lects[1]. Phonotacticon includes the following information about each lect:

- Phonemic inventory (the list of distinctive sound units);
- Tonemes (the list of distinctive tone patterns);
- Onset forms (the list of one or more phonemes that precede the peak of a syllable);
- Nucleus forms (the list of one or more phonemes that form the peak of a syllable); and
- Coda forms (the list of one or more phonemes that follow the peak of a syllable).

While the above information certainly does not cover the entirety of the phonotactic rules of a lect, it does provide comprehensive data on what segments may fill in each of the three slots of a syllable. Phonotacticon allows us to capture phonological diversity that is not only based on the phonemes that are present in each lect, but also based on their distributional characteristics.

The first version of Phonotacticon, or PHONOTACTICON 1.0, covers 516 lects that are spoken in the Eurasian macroarea. In this paper, we will briefly review existing phonological databases (Section 2), explain the construction of Phonotacticon 1.0 (Sections 3-4), and present some descriptive visualizations to indicate what the database can do (Section 5). The paper concludes with suggestions for future research based on Phonotacticon (Section 6).

2 Literature review

As we mentioned, many cross-linguistic phonological databases are currently available, each of them with a different scope of encoded information and a different range of sample lects. Table 1 summarizes seven of the major phonological databases that are currently available. "Size" refers to the number of described lects in each database.

Many quantitative typological studies have been made possible based on these databases. Topics investigated include the areal patterns of consonant phonemes in Eurasia (Nikolaev 2019), the diachronic changes of consonants and vowels across different language families (Moran et al. 2021), and the influence of climate on phonemic inventories (Maddieson & Benedict 2023). A meta-analysis measuring the variance between different databases was also conducted (Anderson et al. 2023), signaling that the databases can significantly differ in the phonological description of the same sample lects, due to the difference in the consulted descriptive works and in how the compilers of each database interpreted those works. This highlights the importance of the coexistence of multiple databases to allow secondary studies to cross-validate their hypotheses using different sources.

Although the number of phonological databases available is growing, there is still a need for a more detailed phonotactic database. While EURPhon (Nikolaev 2018), PBase (Mielke 2008), and LAPSyd (Maddieson et al. 2013) contain different levels of phonotactic information, their information on syllabic phonotactics is relatively limited. In order to fill this gap, Phonotacticon

[1] Available at `zenodo.org/records/10623743`

Name	Size	Containing
PBase (Mielke 2008)	629	Inventory, phonological rules, phonotactics
UCLA Phonological Segment Inventory Database (UPSID, Maddieson 2009)	461	Inventory
Lyon-Albuquerque Phonological Systems Database (LAPSyd, Maddieson 2013a)	683	Inventory, syllable, suprasegmental
The Database of Eurasian Phonological Inventories (EURPhon, Nikolaev 2018)	536	Inventory, phonotactics of Eurasian lects
PHOIBLE 2.0 (Moran & McCloy 2019)	2,186	Inventory
SegBo (Grossman et al. 2020)	574	Borrowed segments
BDPROTO 1.1 (Moran et al. 2021)	257	Inventory of proto- and ancient lects

Table 1: Seven existing phonological databases

1.0 contains detailed information on the possible forms of onset, nucleus, coda in 516 sample lects, providing full segmental information when possible. The next section explains how the 516 lects were sampled.

3 Lect sampling

The 516 sample lects are the lects listed in Glottolog 4.4 (Hammarström et al. 2021), a cross-linguistic bibliographical database, that fulfill the following criteria:

- A living spoken "language" (as defined by Glottolog);[2]
- whose Macroarea is classified as "Eurasia"; and
- whose "Most Extensive Description" as defined by Glottolog is a "long grammar" (i.e. a lect that has at least one lengthy reference grammar published); and
- which had at least one appropriate source accessible to us (see §4.4).

The macroarea "Eurasia" as defined here is the same as the Eurasian continent but excludes most southern Pacific islands typically considered to be part of Eurasia, such as Taiwan or Borneo. This macroarea is defined by Hammarström and Donohue (2014), whose goal was "to come up with a list of objectively predefined areas that can be used as normative controls in cross-linguistic work" (p. 185). Their delimitation of macroareas was purely driven by geographical

[2]Sign lects were not included in the database, as they have distinct phonological systems that cannot be directly compared to spoken phonology.

contiguity (defined by the lack of water body separating landmasses) and not by linguistic genealogy or cultural history. The southern Pacific islands, such as Taiwan, Borneo, or the Philippines, are classified as "Papunesia", except for Hainan, which is separated only by a very thin strait from continental China. Some islands that are too small to be reflected in the resolution of Hammarström and Donohue's study are interpreted as part of a bigger landmass. For example, Ryukyu islands were too small to be reflected in the resolution and were grouped together as the Japanese archipelago, even though some Ryukyu islands are very close to Taiwan.

We limited the sample lects into those having at least one "long grammar" according to Glottolog 4.4. While this choice facilitates the access to the descriptive literature, it biases the samples towards more richly described linguistic families or areas compared to less described ones, as one reviewer pointed out.

The distribution of the 516 sample lects is visualized in Figure 1, where each color-shape combination represents a family. Each of these lects is represented by its phonological profile in the database, which we will explain in the next section.

4 Phonological profile

Phonotacticon consists of the following phonotactic profile of each of the 516 Eurasian lects:

- Phonemic inventory (segmental)
- Tones
- Onset forms
- Nucleus forms
- Coda forms

Table 2 provides an example of the phonological profile of A'ou [aoua1234] (Tai-Kadai; Li et al. 2014). <#> refers to empty onset or coda.

Phoneme	p t k q ʔ pʰ tʰ kʰ qʰ d t͡s t͡ɕ t͡sʰ t͡ɕʰ m n ŋ l ɭ f s ɬ ɕ χ h v z ʑ ɣ ʁ w j a e i ɯ ɔ u ɤ ə o
Tone	55 33 13 31
Onset	p t k q ʔ pʰ tʰ kʰ qʰ d t͡s t͡ɕ t͡sʰ t͡ɕʰ m n ŋ l ɭ f s ɬ ɕ χ h v z ʑ ɣ ʁ w j pl bl vl ml
Nucleus	a e i z ɔ u ai ei ui əu aɯ əɯ iu ɤu ia ie iɔ ua iau iəɯ iəu uai uau uəɯ uei
Coda	# n ŋ

Table 2: Phonological profile of A'ou

How were the five variables chosen? The first two variables, the phonemic inventory and tonemes, are arguably the most basic information of a lect's phonology, as they are present in most of the phonological databases presented in Section 2.

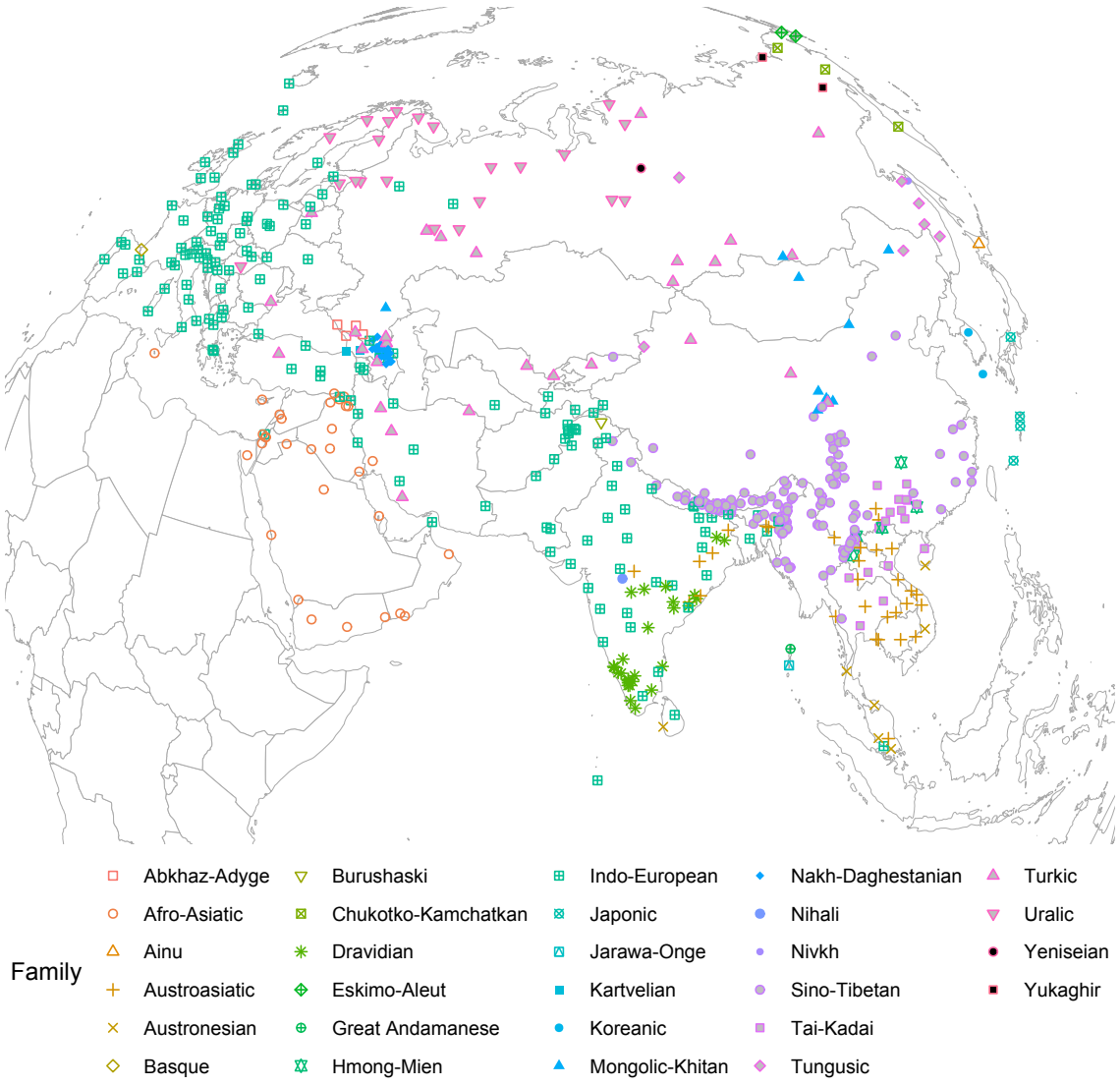

Figure 1: 516 Sample lects of Phonotacticon

The remaining three variables, onset, nucleus, and coda forms, were selected because they form the building units of a SYLLABLE, which is a concept employed by the majority of the phonological analyses of different lects (van der Hulst & Ritter 1999; Goldsmith 2011: cf). An alternative to the syllable would be the notion of WORD, as many phonological analyses describe a lect's phonotactic patterns based on word boundaries, such as word-initial or word-final consonant clusters, rather than syllable boundaries. But previous works on wordhood do not agree on what a phonological or prosodic word is, and many suggest that it is not a cross-linguistically consistent concept (Dixon & Aikhenvald 2003; Schiering et al. 2010). But as the reviewers of this manuscript pointed out, deciding whether to frame the database on syllables or words is ultimately a matter of preference, without any assumption of the theoretical superiority of one over the other. Moreover, when a descriptive grammar provides only word boundary information and no syllable boundary information, we resorted to word boundary information to retrieve syllable boundary information (§4.2).

4.1 Phonemic inventory

The phonemic inventory part of each lect's profile contains the segmental phonemes of the lect. Since Phonotacticon is a phonological database and not a phonetic database, it only lists phonemes as members of the phonemic inventory, excluding its possible allophones.

One challenge of transcribing phonemes using the International Phonetic Alphabet (IPA) is the overspecification of articulatory features. A phoneme is "a combination of certain (simultaneous and/or successive) features, leaving other features unspecified" (Chao 1934: p. 373). The IPA, on the other hand, is an alphabet representing the articulatory possibilities of human speech. As such, every IPA consonant symbol specifies the articulatory place, manner, and phonation of the consonant it represents. The problem arises when a phonologist has to transcribe the nasal phonemes of a lect using the symbols <m n ŋ>, which represent voiced nasal consonants, even when the lect do not specify the feature [+voice] in their nasal phonemes. Thus, using <m n ŋ> to transcribe nasal phonemes that do not specify voicedness is technically an overspecification. Transcribing nasal consonants as <m n ŋ> may be based on the fact that they are articulated as voiced in most environments, but that would be transcribing their (most common) phonetic surface forms and not the combinations of their phonological features, which is the definition of a phoneme. As van der Hulst (2017) puts it, "IPA symbols are mere shorthand for feature representations" (p. 41) and not EQUIVALENT TO the feature representations. Nevertheless, for pragmatic purposes, every phoneme is defined as a IPA symbol in Phonotacticon.

Most of the time, a phoneme is described in the consulted literature as having a single underlying form that can be transcribed as an IPA symbol. But rarely, a phoneme is described as being consisted of several allophones, without a single underlying form. In such case, we judge which allophone represents the underlying form based on which one appears in the most unmarked environments. If there is a form in isolation, then that form is chosen as the representing segment. An example is Japanese moraic nasal /N/, which may occur as [nː], [mː], [ɴː], or others depending on phonotactic context (Iwasaki 2013). /N/ occurs as [ɴː] when it does not precede any segment (e.g. *san* さん [sãɴː] 'three'), so we have transcribed it as /ɴː/.

If there is absolutely no reason to prefer one allophone over another, then we choose the one that appears first in the cited literature. For example, if a phoneme is described as <s/ʃ>, without specifying whether /s/ or /ʃ/ represents underlying form, and there is no reason to believe one is

more unmarked than the other, then it is transcribed in Phonotacticon as <s>.

Archiphonemes - phonemes that have other phonemes as its allophones - are generally treated as equivalent to their allophonic phonemes. Tuvinian [tuvi1240] archiphoneme /I/ can be realized as /i/, /y/, /ɯ/, or /u/, all of which are phonemic in Tuvinian, based on vowel harmony: /àtʰ-I/ > [àtɯ] 'his horse'; /kʰyç-I/ > [kʰyɟy] 'his strength' (Anderson & Harrison 1999: p. 4). In this case, /I/ is treated as equivalent to the phonemes /i y ɯ u/, without being coded as a separate phoneme.

Another problem to be addressed is the XENOPHONE, a phoneme that only occurs in loanwords (coined by Eklund & Lindström 1998). The main complication is that the status of a xenophone can vary from fully nativized to extremely marginal and it can sometimes be difficult to judge whether a xenophone is truly a part of a given lect's phonology. Some xenophones are indistinctively part of a lect's phonology, such as German /ʒ/, while some xenophones are distinctively foreign, such as German nasal vowels, which remain unstable and are often replaced by native phonemes (Wiese 2000: p. 12). Thus, whether a xenophone forms a part of the phonology of a lect is essentially a grey area, leaving us with the question of which xenophones to record in the database and which ones to leave out.

In Phonotacticon, we have included the xenophones as part of the phonemic inventory (and consequently, part of the onset, nucleus, or coda forms) if they are considered to be an integral part of a lect's phonology by the consulted literature. This is mostly inferred from how general a statement is regarding the status of xenophone within a lect's phonology. For instance, if a grammar simply writes "X is a phoneme of this lect" or lists it in within the phonemic inventory table, then we take that to mean that that grammar considers X to be an integral part of the phonemic inventory. On the other hand, if the grammar writes a statement in the lines of "in addition to the above-listed phonemes, X only occurs in some loanwords", then we assume that the grammar does not consider it to be an integral part of the phonology. As ambiguous this strategy of tone-reading can be, it is arguably an appropriate approach to the status of xenophones which is by nature ambiguous. Furthermore, we have excluded xenophones that occur only in certain varieties of the lect and/or freely variable with native phonemes, such as the German nasal vowels.

Phonemes that are used by only a portion of the whole speaker population of a given lect have been excluded as well: we have only included phonemes that are used by all or most speakers. An exception to this rule is that phonemes used by older generations but not by younger generations have been included, due to the fact that younger generations generally reflect the ongoing change of a lect and it is not appropriate to fully reflect an ongoing change as if it were already complete.

In cases where the source describes a sociolinguistic distinction between prescriptive, "educated" speech and real-life, "colloquial" speech, we have generally chosen the latter as better reflecting the phonology of a given lect.

When transcribing the phonemes based on a reference grammar, we relied first and foremost on the articulatory description of that phoneme rather than its orthographic transcription. If a phoneme is transcribed as <c> but described as "voiceless palatal affricate", then we transcribed it as /c͡ç/ (which is the voiceless palatal affricate) rather than the verbatim /c/ (which is the voiceless palatal stop).

All transcribed phonemes are those found in the PanPhon database (Mortensen et al. 2016, as of 23 July 2020). In other words, phonemes that are not found in PanPhon are transcribed in a way that fits PanPhon. The reason for this alignment between Phonotacticon and PanPhon is to make all segments in Phonotacticon machine-readable via PanPhon and convertible into featural

values, hence allowing immediate computational analysis. To align the two databases, some theoretical sacrifices have been made. In the case of diphthongs, PanPhon does not include diphthongs (or triphthongs) as independent segments, even though some grammars argue that a diphthong forms an independent phoneme in the described lect. Hence, even if a diphthong phoneme of a lect consists of two vowels that are not found as monophthongs in that lect, those two vowels nevertheless occur in Phonotacticon as individual phonemes, contrary to the grammar's description. For example, if a grammar describes a lect as having /ɛ͡ɪ/ as a diphthong phoneme while not having /ɛ/ or /ɪ/ as monophthong phonemes, we have still listed /ɛ/ and /ɪ/ as phonemes instead of /ɛ͡ɪ/.

This approach is beneficial to the database, since it not only allows it to be compatible with PanPhon, but also because it avoids the highly controversial nature of the status of diphthongs as individual phonemes. For example, whether diphthongs in a given lect constitute individual phonemes or are combinations of two vowel phonemes is a matter of debate (Pike 1947; Berg 1986; Eliasson 2022) and is thus highly subject to theoretical bias. Another problem is that most reference grammars do not explicitly address this question. Very few grammars –in our experience –discuss whether a diphthong in a lect has phonemic status or not. Many descriptive grammars simply provide a list of monophthongs and diphthongs without specifying whether the diphthongs are sequences of phonemes or independent phonemes. By listing all diphthongs as combinations of monophthong phonemes, we can make all the vowel phonemes compatible with PanPhon and allow cross-linguistic analysis, albeit at the sacrifice of favoring one theoretical approach to diphthongs over another. Moreover, regardless of the phonemic status of diphthongs and triphthongs, they are still listed in the nucleus part of the database, so there is no sacrifice at the descriptive level.

Exceptionally, we have made the following changes to PanPhon:[3]

- The features [hitone] and [hireg] were excluded, since they only pertain to tones and not segments.
- We included prenasalized and preaspirated segments, as these concepts are employed by quite a few grammars but absent in PanPhon. Their features are identical to the nasal and aspirated equivalents, except that prenasalized segments are assigned 0 value to the [nasal, sonorant] features and preaspirated segments are assigned 0 value to the [constricted glottis] feature. The prenasalized consonants are transcribed with <ⁿ> followed by a segment (<ⁿb>, <ⁿd>), whereas preaspirated consonants are transcribed as <h> followed by a tie bar and a segment (<h͡p>, <h͡t>).
- We included the FORTIS (or TENSE) counterpart of all consonants, transcribed by the segmented followed by a small plus sign, as this concept is employed in works on Korean (Lee 2021), Swiss German (Fleischer & Schmid 2006), or other lects but not present in PanPhon. The feature of each fortis consonant is identical to its non-fortis counterpart, except that its [tense] feature is 1 and not 0.
- Some segments that we judge to be missing as accidental gaps were added. For example, /t͡sʼʷː/ was absent in PanPhon, even though /t͡sʼː/ and /t͡sʼʷ/ were present. As such cases are clearly gaps created by mistake, we added such segments in with appropriate feature values.

[3]The revised version of PanPhon is available at `zenodo.org/records/10623743`.

As different phonological databases offer different featural parameters and values per segment, users may convert the IPA segments of Phonotacticon into featural values using databases other than PanPhon, such as PHOIBLE (Moran & McCloy 2019) or the Cross-Linguistic Transcription Systems (Rubehn et al. 2024), although they may not cover every segment in Phonotacticon.

In some cases, a source may specify only a certain class of segments as part of a permissible sequence of phonemes. For example, the source may indicate that a plosive plus a liquid may form an onset cluster, without specifying whether all logically possible combinations of plosive + liquid are permitted in the onset position. In such cases, we have used capital letters to describe the permitted sequence without specifying the segments: PL for plosive (P) plus liquid (L).

Table 3 shows the capital letters used to represent underspecified segments and how they are defined in terms of features and/or graphemes. <j, w, ɥ, ɯɥ> means any segment including any one of these graphemes in its IPA symbol. !<h, ɦ> means any segment not having these graphemes in its IPA symbol. Other than V, which stands for vowels, all the capital letters represent consonants or glides: N refers to nasal consonants and glides only, and excludes nasalized vowels.

Symbol	Class	Features	Graphemes
B	Bilabial	[+cons, +lab]	
C	Consonant	[+cons]	
Č	Affricate	[+cons, +delrel, -son]	
D	Oral	[-nas, -syl]	
F	Fricative	[+cons, +cont, -son]	
G	Glide		<j, w, ɥ, ɯɥ>
K	Coronal	[+cons, +cor]	
Ł	Lateral	[+cons, +cor, +lat]	
L	Liquid	[+cons, +cont, +cor, +son]	
M	Geminate	[+cons]	identical to the previous
N	Nasal	[+nas, -syl]	
P	Plosive	[+cons, -cont, -delrel, -son]	
R	Sonorant	[+cont, +son, -syl]	!<h, ɦ>
S	Sibilant	[+cons, +cont, +cor, -son]	
T	Obstruent	[+cons, -son]	
V	Vowel	[-cons, +cont, +son, +syl]	
W	Voiced	[-syl, +voi]	
X	Voiceless	[-syl, -voi]	
Z	Continuant	[+cont, -syl]	

Table 3: The underspecified segments

Many grammars published in China that describe monosyllabic lects do not describe the lect's phonemic inventory in terms of segmental phonemes but rather in terms of INITIALS (SHENGMU 聲母) and FINALS (YUNMU 韵母), which correspond to onsets and rhymes. When consulting such grammars, we have interpreted the description in terms of phonemes. For example, if a grammar of a lect describes it as having initials /p-, t-, k-/ and finals /-a, -i, -u, -an, -in, -un/, we have interpreted that as a phonemic inventory of /p, t, k, n, a, i, u/.

All geminates are considered to be consonant sequences and not independent phonemes unless the literature explains why they are independent phonemes.

4.2 Onset, nucleus, and coda forms

The onset, nucleus, and coda sections of Phonotacticon describe the possible onset, nucleus, and coda forms of a given lect. They consist of phonemes listed in the phonemic inventory section, as singleton phonemes or a sequence of phonemes. An exception is the OBLIGATORY EPENTHETIC PHONES, which may not be present in the phonemic inventory section but may be present in the onset, nucleus, or coda sections. For example, Bantawa [bant1281] (Sino-Tibetan) does not have a glottal stop as a phoneme, but does have it as an epenthetic phone to fill in the obligatory onset slot (Doornenbal 2009). In this case, <ʔ> was transcribed in the onset section of Bantawa. Epenthetic phones that are only optionally inserted were not included. The null onset and the null coda are represented as <#> in the onset and the coda sections.

Some grammars list word-initial, word-medial, and word-final consonant clusters instead of consonant clusters in onset and coda position. In such case, we have interpreted the data as follows:

- Word-initial clusters are interpreted as onset clusters.
- Word-final clusters are interpreted as coda clusters.
- Word-medial clusters are interpreted as onset consonants, coda consonants, or the mixture of both. If the grammar does not state the syllable boundary that divides a word-medial cluster, we have located the syllable boundary according to the following principles:
  - If a cluster occurs word-initially or word-finally, then we favored the interpretation that it also exists in a word-medial cluster. For example, if /lp/ occurs word-finally, then the medial cluster /lpt/ would be interpreted as /lp.t/, instead of /l.pt/, given that /pt/ does not occur word-initially.
  - If a medial cluster does not contain sequences that appear as initial clusters or final clusters, then we favored the interpretation that reflects the sonority sequencing principle (Clements 1990). The sonority sequencing principle is here defined as the normative sequence of vowel > glide > liquid > nasal > obstruent in relation to the vicinity to the nucleus. For example, if /lp/ does not occur word-finally and /pt/ does not occur word-initially, then the medial cluster /lpt/ would be interpreted as /lp.t/ rather than /l.pt/, because /Vlp/ reflects the sonority sequencing principle (vowel - liquid - obstruent), whereas /ptV/ does not (obstruent - obstruent - vowel). Not reflecting the sonority sequencing principle is preferable to violating it: For example, /mmp/ would be interpreted as /mm.p/, since /Vmm/ does not reflect but does not violate the sequencing principle (vowel - nasal - nasal), whereas /mpV/ violates it (nasal - obstruent - vowel).
  - If a medial cluster contains both an initial cluster and a final cluster, or if a medial cluster does not contain sequences that appear as onset or coda, and if multiple possible interpretations reflect the sonority sequencing principle, then we resorted to the maximal onset principle (Kahn 1976), favoring complex onsets over complex codas. For

example, if /pl/ is an initial cluster and /lp/ is a final cluster, /lpl/ would be interpreted as /l.pl/, instead of /lp.l/.

- For triconsonantal or longer medial clusters, we applied the maximal onset principle within the length of the initial cluster. For example, for a medial cluster /lpml/, we can divide it into /l.pml/ if a three-consonant cluster is attested word-initially. But if only two-consonant clusters are attested as onset, we can only divide it into /lp.ml/.
- Some works (such as Riad 2013) only list the word-initial and word-final clusters and do not list word-medial clusters. In such cases, we interpreted the word-initial and word-final clusters as the same as onset and coda clusters.

As a reviewer pointed out, this approach of interpreting word-based data into syllabic information, rather than directly listing word-based data into the database, may have compromised the rawness and theory-neutrality of the database. But we argue that such interpretation is an essential part of the database, whose goal is to provide the syllabic phonotactic information of each lect, based on the (often incomplete) data provided by the literature. Additionally, we would like to point out that many grammars do directly provide syllabic information. It is only when a grammar did not do this that we had to resort to this interpretation algorithm.

In some cases, a given set of phonemes may be described as permitted in a given position of a sequence. For example, a source may indicate that /p t k s/ may precede /l r w j/ to form a biconsonantal onset cluster, without specifying whether all the 4 * 4 = 16 logically possible combinations are actually attested. In such cases, we have used square brackets to denote ANY ONE OF THE PHONEMES WITHIN THIS BRACKET: [ptks][lrwj] to mean ANY ONE OF /p t k s/ FOLLOWED BY ANY ONE OF /l r w j/.

If a consonant is described as occurring word-initially or as an onset, then we assume that it can occur alone as a single onset. Technically, this may not always be the case, as a consonant may occur word-initially in the onset position as the initial part of a cluster and not on its own (for example, /s/ occurring in /spV/ only and not in /sV/). But unless stated otherwise, we assumed that its occurrence in word-initial or onset position implies its occurrence as a single onset. The same rule applies for word-final and/or coda consonants.

Often, a grammar does not mention whether an onset is obligatory in a syllable. If we detected at least one syllable without an onset, then we judged that that lect does not oblige an onset.

If the literature does not mention syllabic consonants, then we assumed that the syllable requires at least one vowel.

4.2.1 Allophonic variation

A phoneme is only listed at a position of a syllable when it is distinctive in that position, i.e. not neutralized with another phoneme. For example, Korean /t/ and /s/ neutralizes in coda position as [t̚]. One could say that the Korean /s/ is present in coda position, realized as its allophone [t̚]. But because it is not distinctive with /t/ in that position and [t̚] is phonetically closer to [t] than it is to [s], we have listed /t/ as a possible Korean coda but not /s/.

4.2.2 Other rules on segmental transcription

- Dental consonants are transcribed with the dental diacritic (e.g. /t̪ d̪/) only when it is minimally contrastive with alveolar correspondents. Otherwise they are transcribed without

the dental diacritic (e.g. /t̪ d̪/).

- Quite often, <r> is presented as a "liquid" consonant without any specification about its manner or place of articulation. In the absence of additional details, we have transcribed it as /r/.
- The two vowel symbols <ɿ> and <ʅ> that frequently appear in grammars written in China have been interpreted as syllabic consonants /z̩/ and /ʐ̩/, respectively.
- The alveolo-palatal nasal, transcribed as <ȵ> in grammars written in China, are transcribed as the palatal nasal <ɲ> unless it is contrastive with the palatal nasal.
- Some grammars (e.g. Gowda 1968) treat vowel nasalization as a suprasegmental phoneme rather than treating nasal vowels as phonemes. For theoretical consistency, we have interpreted all such cases as independent nasal vowel phonemes.
- Often, a source describes a diphthong as a VV or a GV/VG sequence without specifying whether it occurs within the nucleus or crosses the onset-nucleus or nucleus-coda boundary. Unless stated otherwise, we have assumed that the segments transcribed as vowels, such as /i/ in /ia/ or /ai/, occur within the nucleus, while the segments transcribed as glides, such as /j/ in /ja/ or /aj/, occur in onset or coda position.
- Arabic "emphatic" consonants are transcribed as pharyngealized (<C^{ˤ}>) unless specified otherwise.
- Voiced aspirated obstruents (/b^{h} d^{h} g^{h} .../) are transcribed as breathy obstruents (/b̤ d̤ g̤ .../).

4.3 Tonemes

Tones are transcribed in capital letters (H, M, L, F, R, or any combination of these) or Chao letters (1 to 5 or any combination of these). For example, a high rising tone may be transcribed as HR in capital letters or 35 in Chao letters. If a grammar employs Chao letters, then the Chao letters are transcribed verbatim in Phonotacticon. If a grammar uses other means of description, then the tones are transcribed in capital letters. If a lect has no tones, then the absence of tones is marked with <->.

As a rule, the tones are transcribed in terms of pitch (level or contour) unless a toneme is not distinguishable by pitch only. A toneme often has acoustic cues other than pitch, such as length and phonation. Only when two tonemes are only distinguished by non-pitch cues have we transcribed the non-pitch information in Phonotacticon: <^{ʔ}> for creaky voice, <C> for checked tones, and <^{h}> for aspiration. For example, Burmese tones are transcribed as L (low), H^{ʔ} (high creaky), and H^{h} (high aspirated) (based on Jenny & Hnin Tun 2016).

In some cases, a tone may be described as having more than one allotones, rather than one single underlying form. In those cases, the allotones are transcribed and separated by slashes. For example, the three tones of Asho Chin [asho1236] are transcribed as <55, 44, 22/11> (based on Zakaria 2018).

Many grammars of atonal lects do not specifically mention the absence of tone. If the cited literature does not mention tone, then we have assumed that the lect has no tone.

4.4 Bibliographical sources

The database includes the bibliographical information of the source consulted for each lect's profile. The sources are either the "long grammars" as defined by Glottolog 4.4 or any other source we deem relevant and accessible. As most reference grammars are very long, we did not read each grammar in its entirety (including syntax, morphology and other sections), but rather focused on its phonology section. The consulted phonological data were then transcribed manually into Phonotacticon.

The accessibility issue includes language barriers as well. In most cases, including when the sources were written in French, German, Japanese, or Chinese, this was not a concern, as we could read those lects. In some cases when we could not read the lect a source was written in very well, such as Russian or Finnish, we read it with the aid of machine translation.

4.5 Note

In cases where further clarification is needed regarding how we retrieved the information from the cited source, we left a brief note in plain words in addition to the phonotactic profile.

4.6 Missing information

All the sample lects have, at the very least, its phonemic inventory transcribed. In some cases, the only accessible source did not provide any information on onsets/nuclei/codas. In case where this information was missing, we filled the slot with the symbol <?>. A total of 59 out of the 516 sample lects have no information about its onsets/nuclei/codas.

5 Descriptive visualizations

So far, we have introduced how Phonotacticon 1.0 has been developed. Considering that it is the first database containing the possible onset, nucleus, and coda forms of a sizeable number of lects, we would like to present the possibilities this database can bring. In the following sections, we introduce some visualizations generated from Phonotacticon and discuss areal patterns observable from them.

5.1 Syllable length

In this section, we visualize the distribution of SYLLABLE LENGTH in Eurasia. By syllable length we mean the number of segments (phonemes or epenthetic phones) that fill in the one of these three slots. For example, English permits up to three consonants in its onset position (/**str**ɪt/ *street*, /**spl**æʃ/ *splash*), three vowels in its nucleus position (/f**aɪə**/ *fire*, /**aʊə**/ *hour*), and four consonants in its coda position (/te**ksts**/ *texts*, /glɪ**mpst**/ *glimpsed*) (Gut 2009). English, and European lects in general allow longer onsets, nuclei, and codas compared to other lects in the world. Hokkaido Ainu, for instance, allows only one segment in each of the three positions, its maximal syllable being CVC (Tamura 2000: p. 21).

To our knowledge, Maddieson (2013b) is the only work so far to have provided a typological overview on syllable length. Maddieson divides 486 lects worldwide into three categories based

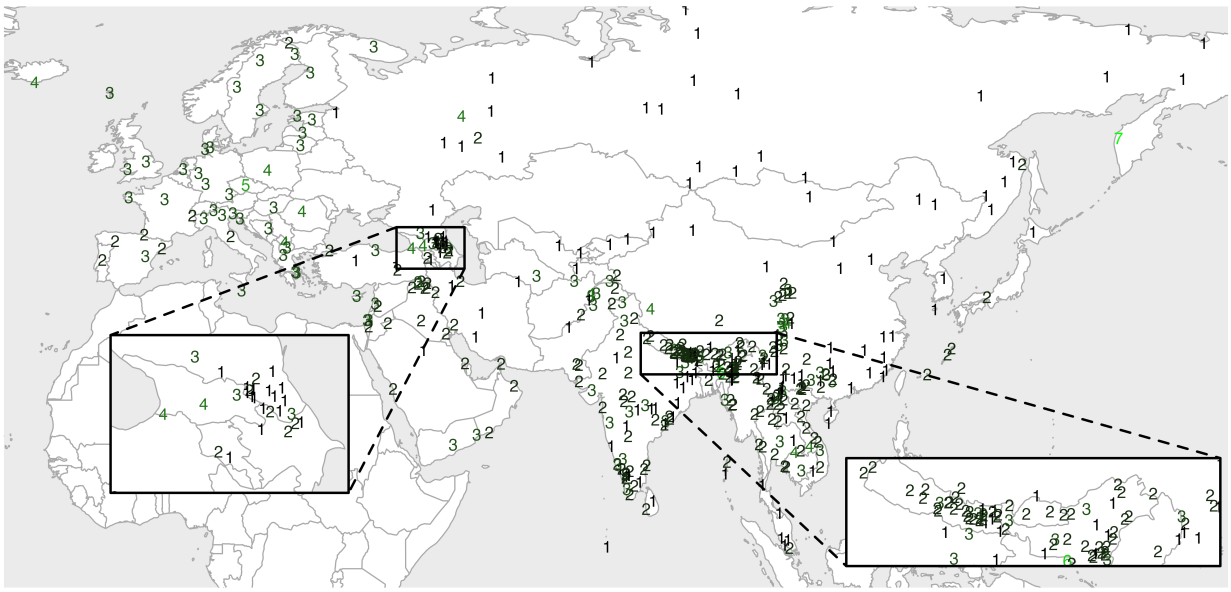

Figure 2: Maximal length of an onset in each lect

on their syllabic complexity: SIMPLE (maximal syllable is CV), MODERATELY COMPLEX (maximal syllable is CCVC where the onset CC is stop + glide or stop + liquid), and COMPLEX (onsets can be clusters other than stop + glide/liquid and codas can be complex). He reports that ca. 56.% of the sample lects have a moderately complex syllable structure, ca. 30.9% have a complex syllable structure, and ca. 12.5% have a simple syllable structure. His data show that within Eurasia, East and Southeast Asian lects tend to allow moderately complex syllable structures, whereas complex syllable structures dominate elsewhere.

Maddieson's overview based on a ternary division based on syllable length, while by itself helpful, calls for a further analysis with finer resolution. The following figures provide such an analysis based on gradient values of onset, nucleus, and coda lengths.

Figure 2 shows the maximal length of an onset in the sample lects, in terms of the number of the phonemes allowed. Caucasus and Eastern Himalayas are zoomed in, due to the high concentration of lects. What is the most evident is that Eurasia is largely divided into three areas: North and Northeast Asia generally only permit singleton onsets, with the notable exception of the Qinghai-Gansu linguistic Area (Janhunen 2006; Dwyer 2013; Xu 2017; Zhou 2020); South and Southeast Asia generally permit up to bisegmental onsets; and Europe generally permits up to triconsonantal onsets. The Middle East seems to be the most diverse without a dominant upper limit.

As the onset is optional in some lects, the minimal length of onset can be either zero or one segment in a given lect. Figure 3 shows whether an onset is obligatory in each lect. We see that the the obligatory onset is mostly present in the Mainland Southeast Asian linguistic area (Enfield 2018; Vittrant & Watkins 2019; Sidwell & Jenny 2021) and the Middle East. All sample lects that mandate an onset in a syllable use the glottal stop [ʔ] as the filler segment to fill in the gap of a syllable that would otherwise lack an onset. The glottal stop may or may not be a phoneme in such lects.

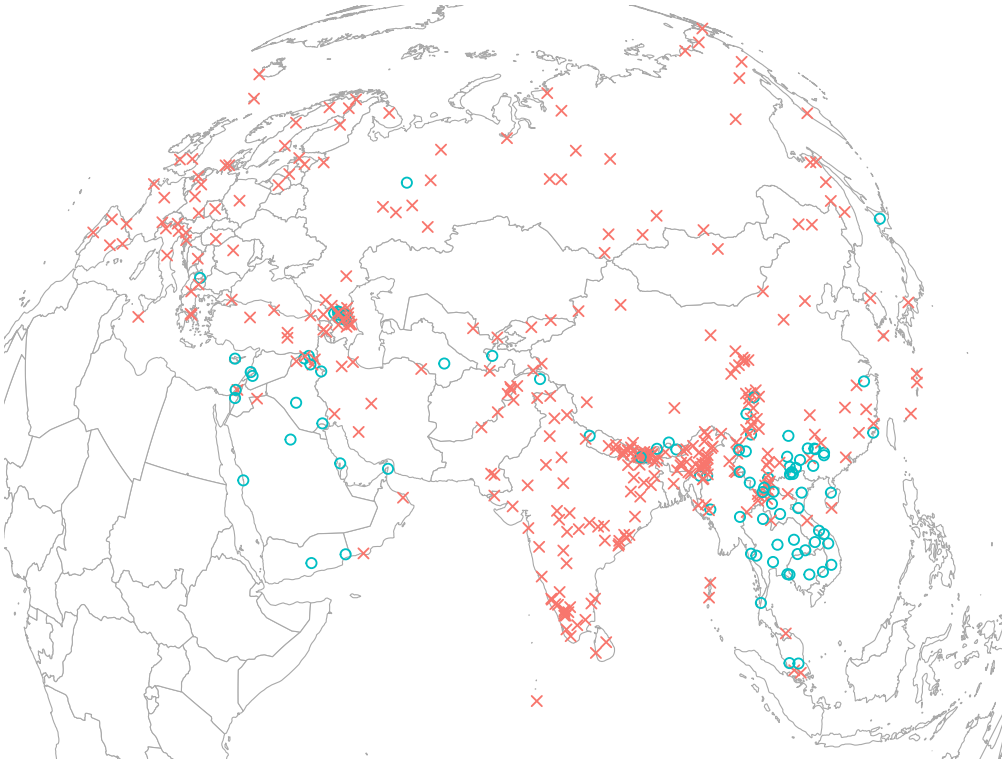

Figure 3: Obligatory onset

Note that even the lects that do not have an obligatory onset filler may have a non-obligatory filler. English, for example, can insert /ʔ/ in the word-initial position, but it is certainly not obligatory (occurring about 50% of the time in British English, according to Fuchs 2015). Furthermore, the glottal stop is normally not inserted in word-medial onsets (e.g. *A. I.* [(ʔ)ɛɪ.ɑɪ] and not *[(ʔ)ɛɪ.ʔɑɪ]).

Figure 4 shows the maximal length of nucleus in each of the Eurasian lects. We see that South, Southwest, and Central Asia tend to not allow complex nuclei, whereas in other areas, diphthongs or even triphthongs are common. Note that lects that only permit monosegmental nuclei may also have phonetic diphthongs if glides appear in their onset or coda position. For instance, according to Bauer and Benedict (1997: p. 57), Cantonese diphthongs are not analyzed as vocalic sequences within a nucleus but rather as a vocalic nucleus followed by a consonantal coda, based on the short duration of the offglides [j] and [w]. Adding to this argument, we can also argue for the nucleus-external hypothesis based on phonological grounds: if the Cantonese diphthongs were nucleus-internal, then it would be difficult to explain why they are not followed by a coda (i.e. *[VVC]). Given the fact that Cantonese only allows one segment as a coda, the impossibility of an offglide and the coda consonant coexisting favors the explanation that an offglide is a coda itself.

Figure 5 shows the maximal length of a coda in each lect. The distribution is very similar to the distribution of maximal onset length shown in Figure 2: European lects allow multiple (as long as six) codas, Southwest and South Asian lects allow up to two, and East Asian lects allow only one. The main difference between onset length and coda length distributions is that Southeast Asian lects do not allow complex codas and that several lects in Southwest China do not allow any coda at all. In sum, we observe a general correlation between onset length and coda length

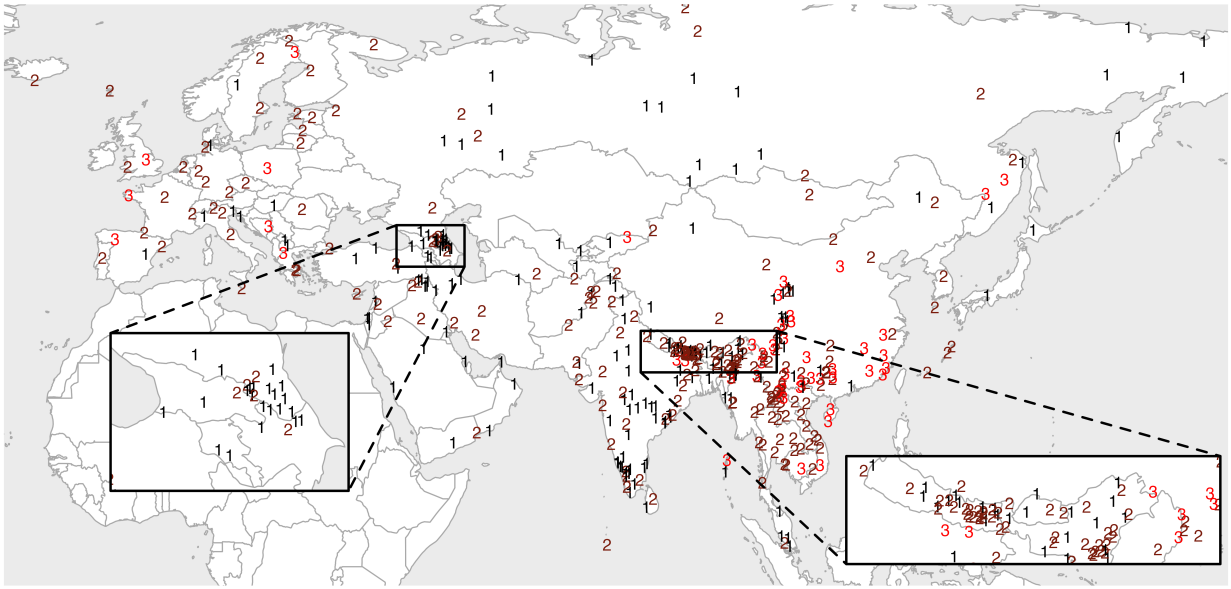

Figure 4: Maximal length of a nucleus in each lect

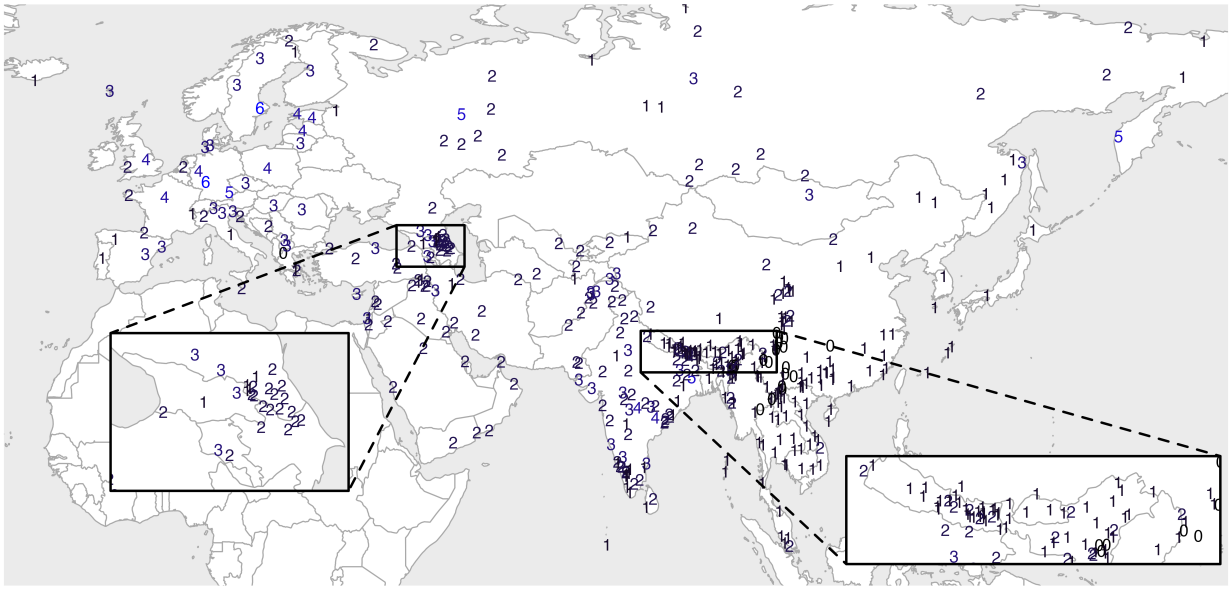

Figure 5: Maximal length of a coda in each lect

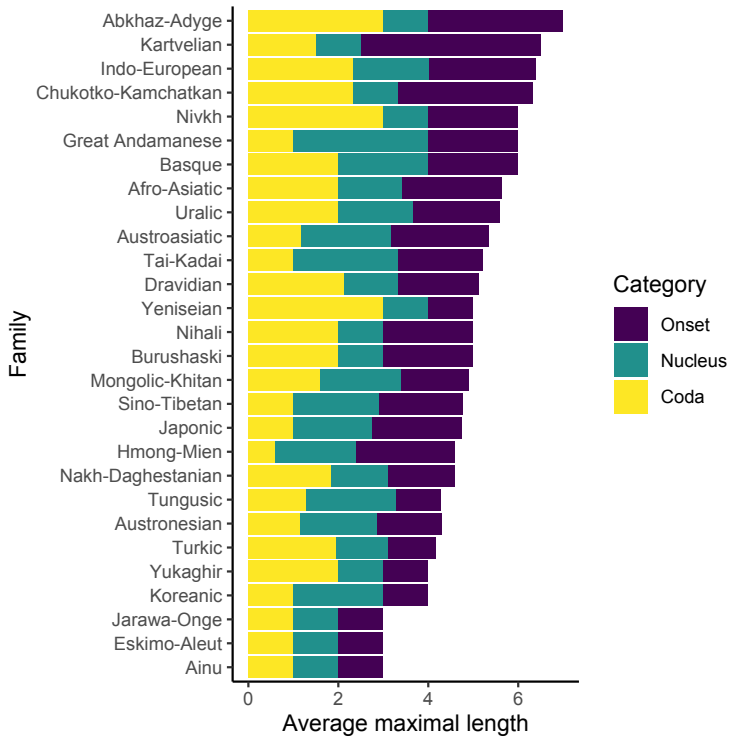

Figure 6: Average maximal length of onset/nucleus/coda by family

in the Eurasian macroarea, which both tend to increase westwards.

Figure 6 shows the average maximal length of onset, nucleus, and coda per each family. We see that generally, language families in western Eurasia, such as Indo-European and Afro-Asiatic, allow more segments per syllable than language families in eastern Eurasia, such as Tungusic and Sino-Tibetan.

To confirm our visual observation that the maximal lengths of onset and coda tend to be longer in western Eurasia compared to eastern Eurasia, we tested whether the maximal onset and coda lengths are correlated with the longitude of the Eurasian lects. First, it is necessary to test the spatial autocorrelation, as geographically neighboring lects may have similar phonotactic patterns. We identified the geographical neighbors of each lect, defined by lects whose coordinates are within 1,500km distance of each other. This distance threshold leaves no sample lect without any neighbor. We then created a weight matrix and assigned the value of 1 to each neighboring lect pairs and the value of 0 to each non-neighboring lect pairs. We also need to test the genealogical autocorrelation, since lects belonging to the same family may have similar onset or coda lengths. Similarly to the spatial weight matrix, we created a genealogical weight matrix where the lect pairs belonging to the same family were assigned the value of 1. Based on these two weight matrices (spatial and genealogical), we performed the Moran's I test to test the spatial and genealogical autocorrelations on onset and coda lengths, which confirms that both onset length and coda length are areally and genealogically clustered ($p < 0.001$).

We then added these two weight matrices together. Thus, lects that are geographical neighbors and also belong to the same family are given the score of 2; lects that are geographical neighbors but belong to different families are given 1; lects that belong to the same family but are

not geographical neighbors are given 1; and the rest are given 0. As a reviewer pointed out, this is under the assumption that spatial autocorrelation and genealogical autocorrelation are of equal weight. The rows of the converged weight matrix were then standardized so that the values of each row add up to 1.

Based on this spatio-genealogical weight matrix, we then ran the Lagrange multiplier diagnostics on the simple linear regressions between longitude and onset/coda lengths to decide whether the spatial lag model or the spatial error model is adequate to perform the spatial regression. The results show that both models are significant. We chose the spatial lag model from a theoretical perspective, since it is more likely that the onset and coda lengths of a lect are influenced by the onset and coda lengths of neighboring lects (spatial lag model) than there are unobserved language-external factors causing onset and coda lengths of neighboring lects to be similar (spatial error model). Finally, based on the spatial lag model, we performed spatial regression to test the correlation between longitude of the lects and their onset/coda length, also weighing in the spatio-genealogical weight matrix. The results show that both onset and coda lengths are correlated with longitude ($p < 0.01$). The likelihood-ratio tests in both regressions show that the rho is significant ($p < 0.001$), confirming that the spatial lag model was suitable. This confirms our visual observation that the maximal onset length and the maximal coda length grow as one goes westwards in Eurasia, even when controlling for spatial and genealogical autocorrelations.

5.2 Syllabic consonants

In all the sample lects, and perhaps universally, the minimal nucleus length is one segment, as a syllable by definition requires at least one segment to form its nucleus. Some lects, however, do not require a vowel in nucleus position, as they allow consonants to form the nucleus. Consonants that form the nucleus are known as SYLLABIC CONSONANTS.

Figure 7 shows the distribution of lects that allow a syllabic consonant as its nucleus (blue circles) and those that do not (red crosses). We observe that syllabic consonants are generally permitted at the two extremes of Eurasia: In East and Southeast Asia and (to a much lesser degree) in Europe. Although not shown in the visualization, the phonotactic patterns of the syllabic consonants in these two areas also tend to differ. In East and Southeast Asia, syllabic nasals tend to occur as monosegmental syllables, such as Yue Chinese *m⁴* 唔 [m̩^{21}] 'not', and syllabic fricatives tend to occur only after homoorganic fricatives, such as Mandarin Chinese *sì* 四 [sz̩^{51}] 'four'. In European lects, however, syllabic consonants have relatively less phonotactic restriction and can occur after a wider range of onsets, such as English *button* [bʌ.ʔn̩] or German *Vogel* [fo.gl̩] 'bird'.

The permitted syllabic consonants are mostly nasals and sometimes liquids or fricatives. This is an unsurprising result confirming that more sonorant segments tend to appear in the nucleus position.

5.3 Number of singleton codas

In Section 5.1, we saw that the maximal coda length varies across Eurasia. Codas are often limited not only quantitatively, but also qualitatively, as many lects only allow a subset of their phonemes to appear in the coda position. Although many lects also ban certain phonemes from the onset position as well, restriction in the coda position tends to be much stronger. For example, Mandarin Chinese only allows /n ŋ/ as codas, while allowing all consonant phonemes but /z ŋ/ as onsets.

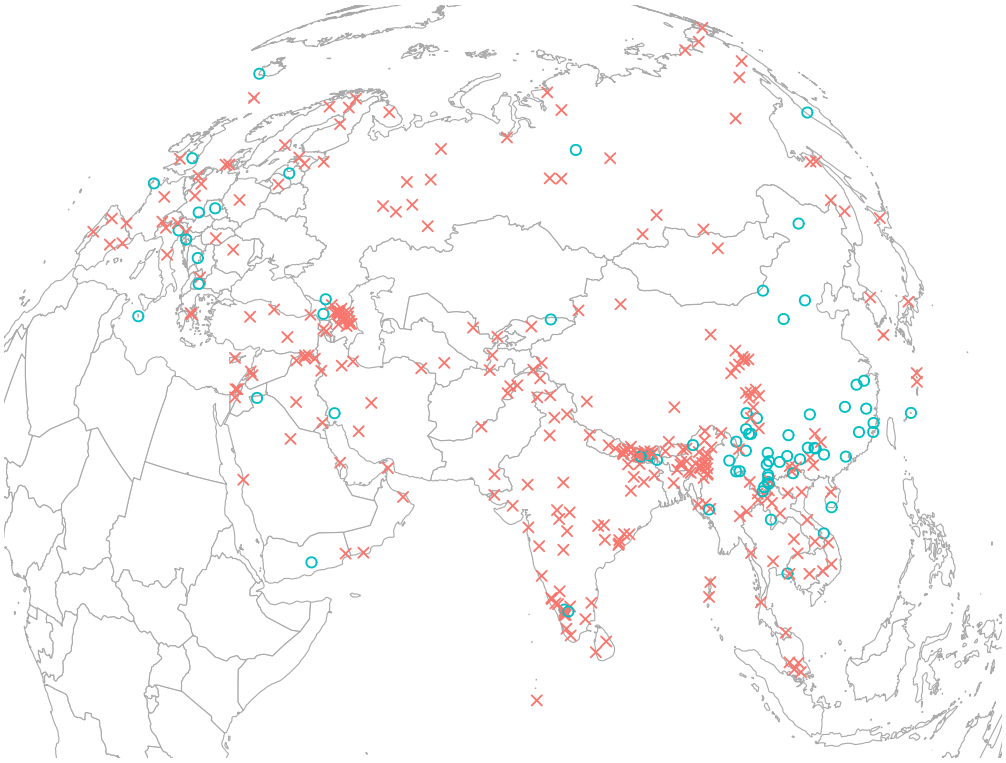

Figure 7: Syllabic consonants

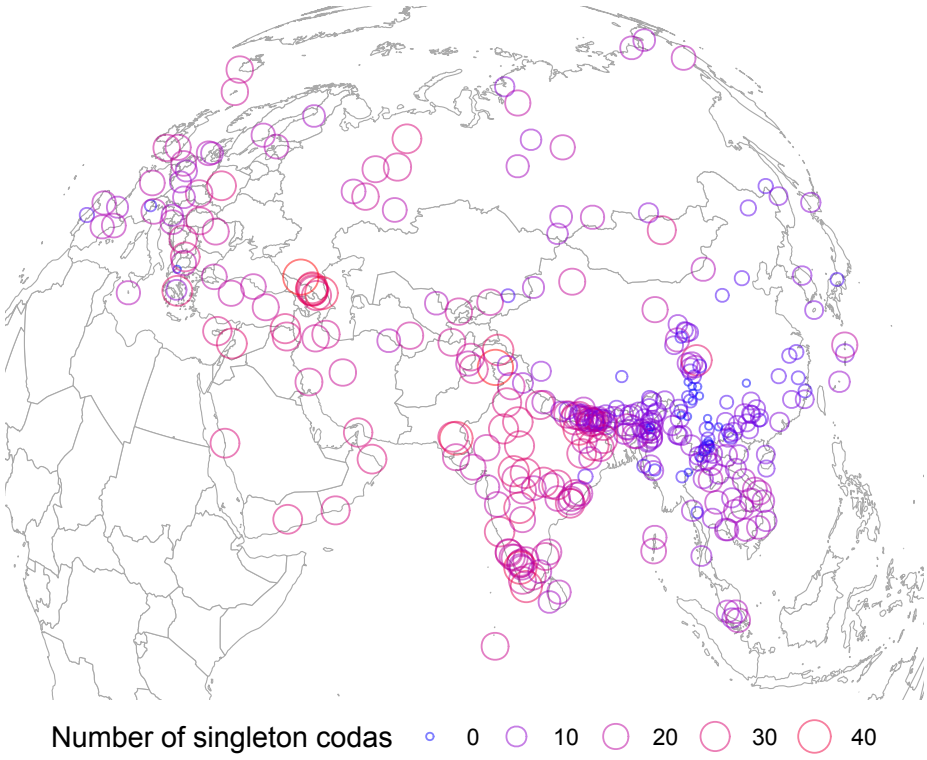

Figure 8: Number of singleton codas

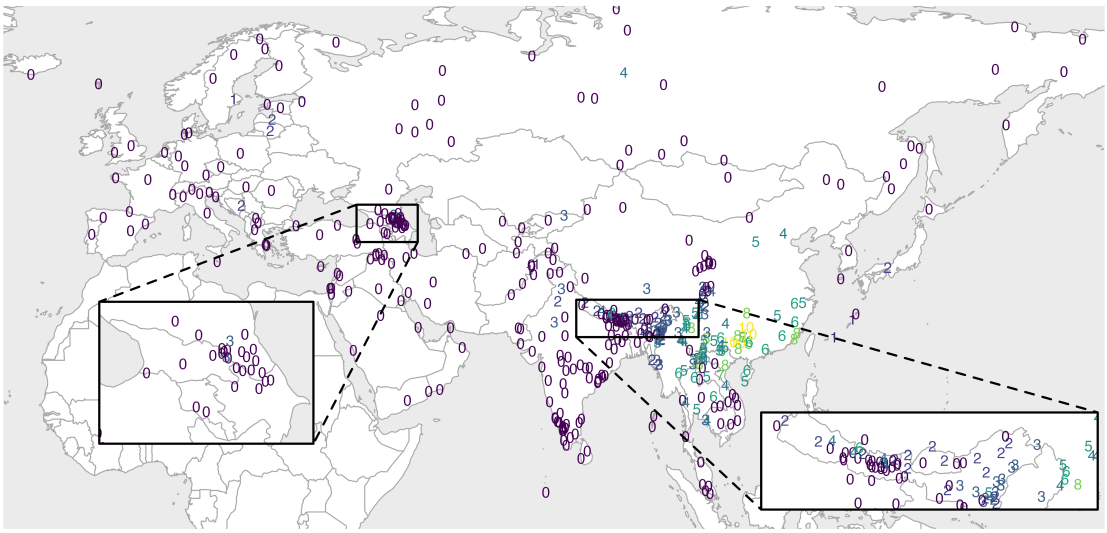

Figure 9: The number of tones per lect

Figure 8 visualizes the types of singleton consonants that can appear as coda, i.e. the types of mono-consonantal codas. (The sample lects are limited to those that have full information of singleton codas, i.e. excluding those whose singleton codas are underspecified as <C> in the database.) It shows that in the lects of East Asia and Southeast Asia, the coda is limited not only in terms of length but also in terms of the number of permitted consonants. Typologically, nasals, plosives and glides are the most common consonants as coda, whereas liquids, fricatives and affricates are less common.

5.4 Number of tones

Maddieson's (2013c) survey of 526 lects worldwide reveals that 220 of them are tonal. Among these tonal lects, 132 have a "simple tone system" with only two tones. The remaining 88 have a "complex tone system" with three or more tones. His data show that tonal lects are most heavily present in Sub-Saharan Africa and Mainland Southeast Asia. Complex tone systems (with three or more tones) are the majority in Mainland Southeast Asia, unlike in Sub-Saharan Africa, New Guinea, or the Americas, where simple tone systems are numerous as well.

Figure 9 shows the number of tones per Eurasian lect. The largest number of distinctive tonemes is ten, for example in Cao Miao (Wu 2015), and the lowest number is one, for example in Swedish, where the tonal distinction is privative, i.e. between the lexical tone and its absence (Riad 2013). It is easily observable that tones are a strongly areal phenomenon, concurring with Maddieson (2013c). Most tonal lects are distributed in Mainland Southeast Asia and China (with the notable exceptions of the Qinghai-Gansu linguistic area, Cambodia, and southern Vietnam). Within this area, the Guangxi province has the highest number of tones, the maximal number being ten. Elsewhere, tones are only sparsely present, lects having at most two tones. From this uneven distribution, we can see that tonogenesis (the emergence of tones) is highly prone to areal pressure, even though it can happen in non-tonal environments (e.g., in Swedish).

It is worth noting that Korean, while depicted as atonal on Figure 9, retains the tones inherited

from Middle Korean in certain varieties (notably the Southeast variety), while the Seoul variety is currently going through tonogenesis (Kang & Han 2013). In the light of the distribution of tones in East Asia, we can hypothesize that Korean tonogenesis may be motivated by areal pressure from Sinitic and Japonic.

5.5 Summary

In this section, we have seen how a number of phonological patterns vary across Eurasia. Crucially, different phonological patterns show different areal distributions: The distribution of tones (Section 5.4), for example, is not identical to the distribution of syllabic consonants (Section 5.2). It is therefore helpful to shed light on each one of the phonological patterns to understand their diverse areal shapes.

6 Conclusion and prospects

In this paper, we have presented the construction of Phonotacticon 1.0, a phonotactic database of Eurasia. In the following years, our goal is to complete Phonotacticon 2.0, a phonotactic database of the world. With some brief descriptive analyses, we have demonstrated that Phonotacticon 1.0 can be a helpful tool for detecting areal phonological patterns across Eurasia. The database further enables us to investigate a variety of interesting topics, including:

- the areality of the phonotactic distributions of certain segments, such as the velar nasal;
- the universality of the sonority sequencing principle, or how well it is observed throughout different lects;
- the correlation between phonotactic parameters, such as between onset length and coda length; or
- which segments most frequently or most rarely appear as codas.

Given the detailed segmental information of a sizeable number of lects and its computational readability via PanPhon, we foresee that Phonotacticon 1.0 and its later versions will inspire a wealth of research into phonological diversity and universality, in Eurasia and beyond.

Data availability statement

Phonotacticon 1.0 and the revised PanPhon are available at:
`zenodo.org/records/10623743`

References

Anderson, Cormac, Tiago Tresoldi, Simon J. Greenhill, Robert Forkel, Russell D Gray & Johann-Mattis List. 2023. Measuring variation in phoneme inventories. *Journal of Language Evolution*. lzad011.

Anderson, Gregory D. & David K. Harrison. 1999. *Tyvan*. München: Lincom.

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

Berg, Thomas. 1986. The monophonematic status of diphthongs revisited. *Phonetica* 43(4). 198–205.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2022. *The AUTOTYP database*. Version v1.1.0. https://doi.org/10.5281/zenodo.6793367.

Blasi, Damián E, Steven Moran, Scott R Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel. 2019. Human sound systems are shaped by post-neolithic changes in bite configuration. *Science* 363(6432). eaav3218.

Blasi, Damián E, Søren Wichmann, Harald Hammarström, Peter F. Stadler & Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* 113(39). 10818–10823.

Chao, Yuen Ren. 1934. The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the National Research Institute of History and Philology* 4(4). 363–398.

Clements, George. 1990. The role of the sonority cycle in core syllabification. In John Kingston & Mary Beckman (eds.), *Papers in laboratory phonology*, vol. 1, 283–333. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511627736.017.

Dixon, R. M. W. & Alexandra Y. Aikhenvald. 2003. Word: A typological framework. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Word: A cross-linguistic typology*, 1–41. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486241.002.

Doornenbal, Marius. 2009. *A grammar of Bantawa: Grammar, paradigm tables, glossary and texts of a Rai language of eastern Nepal*. Rijksuniversiteit te Leiden dissertation.

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.

Eklund, Robert & Anders Lindström. 1998. How to handle "foreign" sounds in Swedish text-to-speech conversion: Approaching the 'xenophone' problem. In *5th International Conference on Spoken Language Processing, 30th November-4th December, 1998, Sydney, Australia*, vol. 7, 2831–2834.

Eliasson, Stig. 2022. The phonological status of Swedish *au* and *eu*: Proposals, evidence, evaluation. *Nordic Journal of Linguistics*. 1–42. https://doi.org/10.1017/s0332586522000233.

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

Fleischer, Jürg & Stephan Schmid. 2006. Zurich German. *Journal of the International Phonetic Association* 36(2). 243–253.

Fuchs, Robert. 2015. Word-initial glottal stop insertion, hiatus resolution and linking in British English. In *Sixteenth Annual Conference of the International Speech Communication Association*, 1675–1679. https://doi.org/10.21437/Interspeech.2015-386.

Goldsmith, John. 2011. The syllable. In John Goldsmith, Jason Riggle & Alan C. L Yu (eds.), *The handbook of phonological theory*, 2nd edn., 164–196. Chichester, West Sussex: Wiley. https://doi.org/10.1002/9781444343069.ch6.

Gowda, K. S. Gurubasave. 1968. *Descriptive analysis of Soliga*. Deccan College dissertation.

Grossman, Eitan, Elad Eisen, Dmitry Nikolaev & Steven Moran. 2020. SegBo: A database of borrowed sounds in the world's languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5316–5322. European Language Resources Association.

Gut, Ulrike. 2009. *Introduction to English phonetics and phonology*. Vol. 1. Frankfurt am Main: Peter Lang.

Hammarström, Harald & Mark Donohue. 2014. Some principles on the use of macro-areas in typological comparison. *Language Dynamics and Change* 4(1). 167–187.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. *Glottolog 4.4*. Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.4761960.

van der Hulst, Harry & Nancy A Ritter. 1999. Theories of the syllable. In Harry van der Hulst & Nancy A Ritter (eds.), *The syllable: Views and facts*, 13–52. Berlin; Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110806793.13.

Iwasaki, Shoichi. 2013. *Japanese*. Revised. Amsterdam: John Benjamins Publishing Company.

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.

Kahn, Daniel. 1976. *Syllable-based generalizations in English phonology*. Massachusetts Institute of Technology dissertation.

Kang, Yoonjung & Sungwoo Han. 2013. Tonogenesis in early contemporary Seoul Korean: a longitudinal case study. *Lingua* 134. 62–74.

Lee 이, Jinho 진호. 2021. *Kwuke umwunlon kanguy* 국어 음운론 강의 *[a course in korean phonology]*. Seoul 서울: Jipmundang 집문당.

Li, Xia, Jinfang Li & Yongxian Luo. 2014. *A grammar of Zoulei (southwest China)*. Bern: Peter Lang.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9(1). https://doi.org/10.1038/s41597-022-01432-0.

Maddieson, Ian. 2009. *Patterns of sounds*. Cambridge: Cambridge University Press.

Maddieson, Ian. 2013a. Consonant inventories. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/feature/1A.

Maddieson, Ian. 2013b. Syllable structure. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/12.

Maddieson, Ian. 2013c. Tone. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/13.

Maddieson, Ian & Karl Benedict. 2023. Demonstrating environmental impacts on the sound structure of languages: challenges and solutions. *Frontiers in Psychology* 14. https://doi.org/10.3389/fpsyg.2023.1200463.

Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Christophe Coupé & François Pellegrino. 2013. LAPSyd: lyon-albuquerque phonological systems database. In *Interspeech 2013*. International Speech Communication Association (ISCA). https://doi.org/10.21437/interspeech.2013-660.

Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press.

Moran, Steven, Eitan Grossman & Annemarie Verkerk. 2021. Investigating diachronic trends in phonological inventories using BDPROTO. *Language Resources and Evaluation* 55(1). 79–103.

Moran, Steven & Daniel McCloy. 2019. *Phoible 2.0*. Max Planck Institute for the Science of Human History. https://phoible.org/.

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.

Nikolaev, Dmitry. 2018. The database of Eurasian phonological inventories: A research tool for distributional phonological typology. *Linguistics Vanguard* 4(1).

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

Pike, Kenneth L. 1947. On the phonemic status of english diphthongs. *Language* 23(2). 151–159.

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

Rubehn, Arne, Jessica Nieder, Robert Forkel & Johann-Mattis List. 2024. Generating feature vectors from phonetic transcriptions in cross-linguistic data formats. *arXiv preprint arXiv:2405.04271*. https://doi.org/10.48550/arXiv.2405.04271.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7(1). https://doi.org/10.1038/s41597-019-0341-x.

Schiering, René, Balthasar Bickel & Kristine A. Hildebrandt. 2010. The prosodic word is not universal, but emergent. *Journal of Linguistics* 46(3). 657–709.

Sidwell, Paul & Mathias Jenny. 2021. *The languages and linguistics of Mainland Southeast Asia: A comprehensive guide*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110558142.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175. https://doi.org/10.1126/sciadv.adg6175.

Tamura, Suzuko. 2000. *The Ainu language*. 1st edn. Tokyo: Sanseido.

van der Hulst, Harry. 2017. Phonological typology. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *The Cambridge handbook of linguistic typology*, 39–77. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316135716.002.

Vittrant, Alice & Justin Watkins (eds.). 2019. *The Mainland Southeast Asia linguistic area*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110401981.

Wiese, Richard. 2000. *The phonology of German*. Oxford: Oxford University Press.

Wu, Manxiang. 2015. *A grammar of Sanjiang Kam*. University of Hong Kong dissertation.

Xu, Dan. 2017. *The Tangwang language: An interdisciplinary case study in northwest China*. Cham: Springer.

Zakaria, Muhammad. 2018. *A grammar of Hyow*. Nanyang Technological University dissertation.

Zhou, Chenlei. 2020. Case markers and language contact in the Gansu-Qinghai linguistic area. *Asian Languages and Linguistics* 1(1). 168–203.