Phonotacticon: A cross-linguistic phonotactic database

Ian Joo[1,2*] and Yu-Yin Hsu[2]

[1]Faculty of International Studies, Nagoya University of Commerce and Business, 4-4 Sagamine, Komenoki-cho, Nisshin-shi, 470-0193, Aichi, Japan.
[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, 11 Yuk Choi Rd., Hung Hom, Kowloon, Hong Kong SAR.

*Corresponding author(s). E-mail(s): ian_joo@nucba.ac.jp;
Contributing authors: yu-yin.hsu@polyu.edu.hk;

Abstract

Phonotacticon is a cross-linguistic database that contains basic 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 most of the spoken lects that are spoken across the world. 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.

Keywords: phonotactics; phonology; database; typology

1 Introduction

Constructing quantitative typological databases may be one of the first crucial steps in enabling research on computational 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 and 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).

Perhaps 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. (In this article, we use the term *lect* to refer to any level of linguistic variety, commonly referred to as a *dialect* or a *language*, as the distinction between a dialect or a language is sociocultural in nature and thus not directly relevant to typological research). 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 phonotactic information of spoken lects, which is available at github.com/ianjoo/Phonotacticon. 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 the 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 the existing phonological databases (Section 2), explain the construction of Phonotacticon 1.0 (Sections 3-5), and present some descriptive visualizations to indicate what the database can do (Section 6). The paper concludes with suggestions for future research based on Phonotacticon (Section 7).

2 Literature review

In this section, we review eight of the most important phonological databases, and focus on those that are currently accessible.

2.1 UCLA Phonological Segment Inventory Database (UPSID)

The UCLA Phonological Segment Inventory Database, or UPSID (Maddieson, 2009), which is accessible at web.phonetik.uni-frankfurt.de/upsid.html), and was released in 1984, is the oldest phonological database that is currently available online. It consists of the phonemic inventory of 451 lects across the world. Although it is not without limitations, such as only containing segmental information and not tones, UPSID remains a useful phonological database at present.

2.2 The Database of Eurasian Phonological Inventories (EURPhon)

The Database of Eurasian Phonological Inventories, or EURPhon (Nikolaev, 2018), which is accessible at eurphon.info, describes the phonological inventories of 536 Eurasian lects. It also contains some phonotactic information for many of the lects, such as word-initial consonant clusters, word-final consonants, and possible syllabic templates. This database is possibly the database that is the most similar to Phonotacticon 1.0, which also provides the phonotactic profiles of Eurasian lects, even though the two databases bear some structural differences, as will be explained in Section 5.

2.3 PHOIBLE 2.0

PHOIBLE 2.0 (Moran and McCloy, 2019), which is accessible at phoible.org, may be the largest and the most widely used phonological database at present. Similar to UPSID but on a much larger scale, PHOIBLE 2.0 contains the phonological inventories of 2,186 lects worldwide. One of its strengths is that it often includes multiple inventories for each lect retrieved from different sources (including UPSID and EURPhon), thus, enabling cross-doculect comparisons. For instance, four inventories are available for Korean. This is quite useful considering that different descriptions of a lect's phonological inventory can vary to a significant degree depending on the consulted bibliographical source (Anderson et al, 2021). Unlike UPSID, it also describes the tonemes of the tonal languages.

2.4 PBase

PBase (Mielke, 2008), which is accessible at pbase.phon.chass.ncsu.edu, provides the following phonological information for each of the 629 lects:

- Core inventory
- Marginal inventory
- Phonotactic distribution
- Phonological rules

As an example, for Indonesian (pbase.phon.chass.ncsu.edu/language/4), PBase lists /p t ʧ k ʔ b d ʤ g s h i u e ə o m n ɲ ŋ a l r w j/ as its core inventory and /f ʃ x z/ as its marginal inventory (in this case, xenophones). It provides non-exhaustive information about its phonotactic distribution, such as only /p t k ʔ s h m n ŋ l r w j/ appearing as the morpheme-final consonant. It also provides a non-exhaustive list of phonological rules, such as /p t k/ being unreleased word-finally.

To our knowledge, PBase is the only phonological database that distinguishes the marginal inventory from the core inventory and provides phonological rules, such as allophonic variations. Although the marginality of phonemes in any lect is a continuous feature rather than a categorical one, with some phonemes being less marginal than others, it is nevertheless extremely useful to have a binary distinction between marginal and core inventories, as the two inventories often behave differently in phonotactic terms. Phonological rules can also provide useful information about cross-linguistic phonological patterns, as many phonological rules, such as final devoicing, are shared by different lects.

However, the highly uneven distribution of the phonological information about different lects makes PBase less suitable for quantitative cross-linguistic comparisons. For example, English has 36 rules and distributions coded in the database, whereas Ainu only has seven. Phonotacticon may overcome this problem by having a fixed set of variables for each lect (phonemes, tones, onset, nucleus, and coda), although some of the lects in Phonotacticon also lack one or more of these five variables as well.

2.5 Lyon-Albuquerque Phonological Systems Database (LAPSyD)

The *Lyon-Albuquerque phonological systems database* or LAPSyD (Maddieson et al, 2013), which is accessible at lapsyd.huma-num.fr/lapsyd, is a database that is based on UPSID and contains the following phonological information for each of the 683 lects across the world:

- Segmental inventory (including notes on consonants and vowels)
- Diphthongs
- Syllable structures
- Comments on tone and stress
- Location

Perhaps one of the greatest benefits of LAPSyD is its qualitative details, especially for suprasegmental aspects such as stress, which are better explained qualitatively. The great amount of such detailed explanations in a verbal format makes this database extremely useful for a lect-by-lect comparison.

2.6 BDPROTO 1.1

BDPROTO 1.1 (Moran et al, 2021), which is accessible at github.com/bdproto, contains the phonological inventory of 257 ancient and reconstructed lects worldwide. To our knowledge, it is the only phonological database that focuses on non-contemporary lects. As Moran et al. ((Moran, Grossman, and Verkerk, 2021, 87)) pointed out, the time periods of proto-lects are not uniform: Proto-Indo-European was not spoken contemporaneously with Proto-Austronesian. The authors have included the approximate time period for each proto-lect, thus making BDPROTO a useful database for conducting a diachronic analysis of phonological typology.

2.7 SegBo

SegBo (Grossman et al, 2020), which is accessible at github.com/segbo-db), is a list of the borrowed segments in 574 lects worldwide. According to SegBo, /f/ is the most commonly borrowed segment worldwide. As SegBo also codes the donor lect for each segment, it shows that the following five lects are the most prolific donors: Spanish, English, Arabic, Russian, and Indonesian. As segment borrowing is one of the most visible outcomes of language contact, SegBo allows us to detect contact phenomena across the world, especially the asymmetrical contact between less spoken lects and larger, more dominant lects.

One of Segbo's limitations (as described in Grossman et al (2020)) is that it is not areally balanced, as it over-represents certain regions, such as Papunesia and eastern Russia. East Asian lects are relatively underrepresented in the database, hence the underrepresentation of Mandarin Chinese as a donor lect. However, as SegBo is still in the early stages, this problem can easily be overcome by adding more sample lects.

2.8 World Phonotactics Database

The *World Phonotactics Database* (WPD) is a currently inaccessible database compiled by Mark Donohue and his team. It provided phonotactic information of thousands of lects around the world and is perhaps the largest phonotactic database ever published to this day. Personal communication with Mark Donohue and Siva Kalyan (who will conduct analysis using the database) let us know that it will be available online again in the near future.

The database offered data as values of a set of parameters (such as *this lect only allows nasals as codas*) and not as segments (such as *this lect allows /m n ŋ/ as codas)* (Siva Kalyan, personal communication). This is an important distinction between the World Phonotactics Database and Phonotacticon, as Phonotacticon provides every part of the data as segments and tonemes and not as parameter values. This allows us to analyze the phonological distance between lects using a different methodology.

2.9 Summary of phonological databases

Table 1 summarizes the eight databases we have reviewed in this section.

Name	No. of lects	Area	Containing	Available
UPSID	461	World	Inventory	Yes
EURPhon	536	Eurasia	Inventory, phonotactics	Yes
PHOIBLE 2.0	2,186	World	Inventory	Yes
PBase	629	World	Inventory, phonotactics	Yes
LAPSyD	683	World	Inventory, syllable, suprasegmental	Yes
BDPROTO 1.1	257	World	Inventory	Yes
SegBo	574	World	Borrowed segments	Yes
WPD	Thousands	World	Phonotactics	No

Table 1 Summary of the eight databases reviewed

Although the number of phonological databases available is growing, there is still the need for a **form-based phonotactic database**. While EURPhon (Nikolaev, 2018), PBase (Mielke, 2008), and LAPSyD (Maddieson et al, 2013) contain different levels of phonotactic information, they are primarily a database of segmental inventories and their phonotactic information is relatively limited. While we can hope that the World Phonotactics Database will be available again soon, it is a parameter-based database, in which each lect bearing different values of a set of phonological parameters, and not a form-based database containing possible phonological forms a lect can generate according to its phonotactic rules. This form-based phonotactic database is what Phonotacticon is. It contains the basic phonological profiles of lects worldwide, now containing 516 Eurasian lects.

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)[1];
- 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.

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 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.

The distribution of the 516 sample lects is visualized in Figure 1, where each color-shape combination represents a family.

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 (Tai-Kadai; Li et al (2014)).

Phoneme	p t k q ʔ p^{h} t^{h} k^{h} q^{h} d t͡s t͡ɕ t͡s^{h} t͡ɕ^{h} m n ŋ l ɭ f s ɬ ç 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 p^{h} t^{h} k^{h} q^{h} d t͡s t͡ɕ t͡s^{h} t͡ɕ^{h} m n ŋ l ɭ f s ɬ ç 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

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

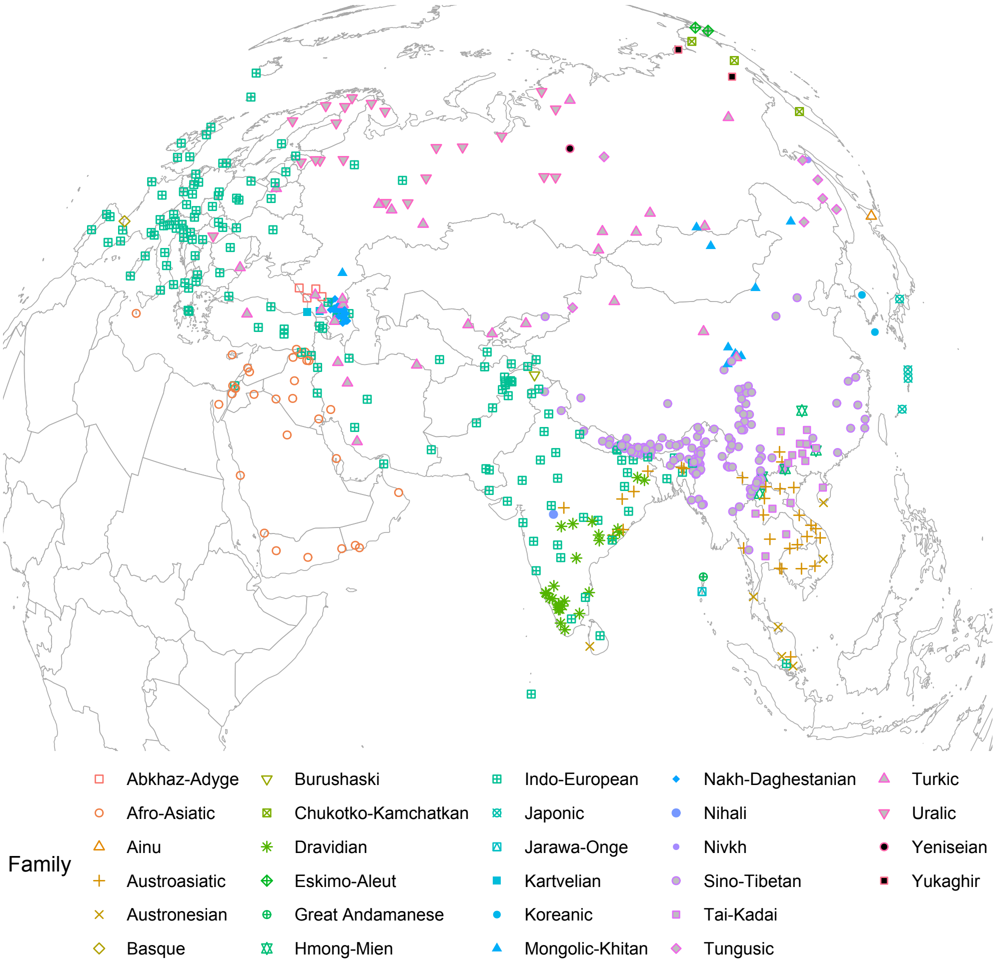

Fig. 1 516 Sample lects of Phonotacticon

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.

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 and Ritter, 1999; Goldsmith, 2011, cf). Malaia and Wilbur (2020) argue that the syllable is a universal strategy to divide continuous linguistic information into discrete segments, observable in spoken and the sign modalities alike. As the syllable is a widely accepted theoretical notion whose reality is supported by neurolinguistic evidence, it is the most suitable framework to be adopted for a cross-theoretical database like Phonotacticon.

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 and Aikhenvald, 2003; Schiering et al, 2010). Thus, it is more cross-theoretically consistent to unify the variables of the database into syllabic notions rather than wordhood notions.

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.

The challenge of transcribing a phonemic inventory using the International Phonetic Alphabet (IPA) is that while a phonemic inventory is a set of combinations of distinctive features, the IPA is an alphabet representing the articulatory possibilities of human speech. For example, the IPA symbol <p> represents the unaspirated voiceless bilabial plosive. While we can use this symbol to represent the English phoneme /p/, which is a bilabial plosive (which can be aspirated or unaspirated based on its environment), using the symbol <p> overspecifies this phoneme in terms of aspiration, as English /p/ can be either aspirated (as in *pan* [p^{h}æn]) or unaspirated (as in *span* [spæn]). In this sense, 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 more than one allophones, without a single underlying form. In such case, any of the allophones are chosen as the underlying form, normally 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/ is the underlying form, then it is transcribed in Phonotacticon as <s>. 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/.

Archiphonemes, phonemes that have other phonemes as its allophones, are generally treated as equivalent to their allophonic phonemes. Tuvinian archiphoneme /I/ can be realized as /i/, /y/, /ɯ/, or /u/, all of which are phonemic in Tuvinian, based on vowel harmony: /àt^{h}-I/ > [àtɯ] 'his horse'; /k^{h}yç-I/ > [k^{h}yʝy] 'his strength' (Anderson and Harrison, 1999, 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. An example is the English /ʒ/, which only occurs in (mostly French) loanwords, such as *measure* and *occasion*. In many cases (but not all), the grammar of a lect mentions that certain phonemes only occur in loanwords.

The "foreignness" of a xenophone is a matter of degree. Some xenophones are indistinctively part of a lect's phonology, such as English /ʒ/, as most native speakers are unaware that it is a loan phoneme at all. At the other end of the spectrum, however, some xenophones are distinctively foreign, such as English /x/, which only occurs in a handful of words like *Bach* or *loch*, and most phonologists would not consider /x/ on par with "native" English phonemes, such as /p/ and /n/. Thus, whether a xenophone forms a part of the phonology of a lect is essentially a grey area.

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. This is mostly inferred from how general a statement is regarding the status of xenophone within a lect's phonology. For instance, if an English grammar simply writes "/ʒ/ is an English phoneme" or lists it in within the phonemic inventory table of English, then we take that to mean that that grammar considers /ʒ/ as an integral part of the English phonemic inventory. On the other hand, if the grammar writes a statement like "in addition to the above-listed English phonemes, /x/ only occurs in some loanwords", then we assume that the grammar does not consider it to be an integral part of English 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. For example, some Korean speakers use [f] in certain loanwords, such as *pail* 파일 [fa.il] '(computer) file' or *polte* 폴더 [fol.dʌ] '(computer) folder'. Importantly, however, (i) not all speakers pronounce these words with [f]; and (ii) it is freely variable with [p^h], which is a native Korean phoneme. Such xenophones were not considered to be an integral part of a lect's phonology.

Phonemes that are used by only a portion of the whole speaker population of a given lect were excluded as well, only including phonemes that are used by all or most speakers. An exception to this rule is that phonemes used by the older generations but not by the younger generations were 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 case where the source describes a sociolinguistic distinction between prescriptive, "educated" speech and real-life, "colloquial" speech, we generally chose the latter as better reflecting the phonology of a given lect.

When transcribing the phonemes based on a reference grammar, we rely 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 transcribe 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. This is especially important for the case of diphthongs, as PanPhon does not include diphthongs (or triphthongs) as independent segments, even though some grammars argues that a diphthong forms an independent phoneme in the described lect. Even if a diphthong phoneme of a lect consists of two vowels that are not found as monophthongs in that lect, those two vowels are nevertheless listed 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 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 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. 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:

- The features [hitone] and [hireg] were excluded, since they only pertain to tones and not segments.
- we have 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 (<^{n}b>, <^{n}d>), whereas preaspirated consonants are transcribed as <h> followed by a tie bar and a segment (<h͡p>, <h͡t>).
- we have 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 Lee 이 (2021)), Swiss German (Fleischer and 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, /t͡s’^{w}ː/ was absent in PanPhon, even though /t͡s’ː/ and /t͡s’^{w}/ were present. As such cases are clearly gaps created by mistake, we added such segments in with appropriate feature values.

The revised version of PanPhon is available at github.com/ianjoo/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 the capital letters to describe the permitted sequence without specifying the segments: PL for plosive (P) plus liquid (L).

Table 3 show the capital letters used to represent underspecified segments and how they are defined in terms of features and/or graphemes. <j, w, ɥ, ɰ> 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, excluding 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, ɰ>
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 will describe the possible onset, nucleus, and coda forms of a given lect. They will 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 (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 interpret 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 locate the syllable boundary according to the following principles:
  - If a cluster occurs word-initially or word-finally, then we favor the interpretation that it also exists in a word-medial cluster. For example, if /lp/ occurs word-finally, then the medial cluster /lpt/ is 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 favor 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/ is 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 preferred to violating it: For example, /mmp/ is 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 resort 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/ is interpreted as /l.pl/, instead of /lp.l/.
  - For triconsonantal or longer medial clusters, we apply 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 interpret the word-initial and word-final clusters as the same as onset and coda clusters.

In some cases, a given set a 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 be always 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 assume 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 detect at least one syllable without an onset, then we judge that that language does not oblige an onset.

If the literature does not mention syllabic consonants, then we assume 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. /t d/).
- 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 transcribe it /r/.
- The two vowel symbols <ɿ> and <ʅ> that frequently appear in grammars written in China are interpreted as syllabic consonants /z̩/ and /ʐ̩/, respectively.
- The alveol-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 assume 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 (<Cˤ>) unless specified otherwise.
- Voiced aspirated obstruents (/bʰ dʰ gʰ .../) 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: <?> 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 and Hnin Tun (2016)).

In some cases, a tone may be described as 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 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 assume 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. The accessibility issue includes language barrier 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 very well the lect a source was written in, 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 have left a brief note in plain words in addition to the phonotactic profile.

5 Difference from EURPhon

Although Phonotacticon 1.0 is similar to EURPhon (Nikolaev, 2018), introduced in Section 2.2, the two databases differ in several regards, namely:

- EURPhon contains the phonotactic constraints on **word boundaries** (word-initial clusters and word finals), whereas Phonotacticon contains the phonotactic constraints on **syllabic components** (onset, nucleus, and coda);
- EURPhon does not contain coda clusters or word-final clusters; and
- EURPhon does not specify syllabic consonants when a sample lect has any.

6 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.

6.1 Syllable length

In this section, we will 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**it/ *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, 21).

To our knowledge, Maddieson (2013a) is the only work so far to have provided a typological overview on syllable length. Maddieson divided 486 lects worldwide into three categories based on their syllabic complexity: *Simple* (maximal syllable is CV), *moderately complex* (maximal syllable is CCVC), and *complex* (maximal syllable is longer than CCVC). 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 shows 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 will 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. 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, cf); South and Southeast Asia generally permit up to bisegmental onsets; and Europe generally permit up to triconsonantal onsets. The Middle East seems to be the most diverse without a dominant upper limit.

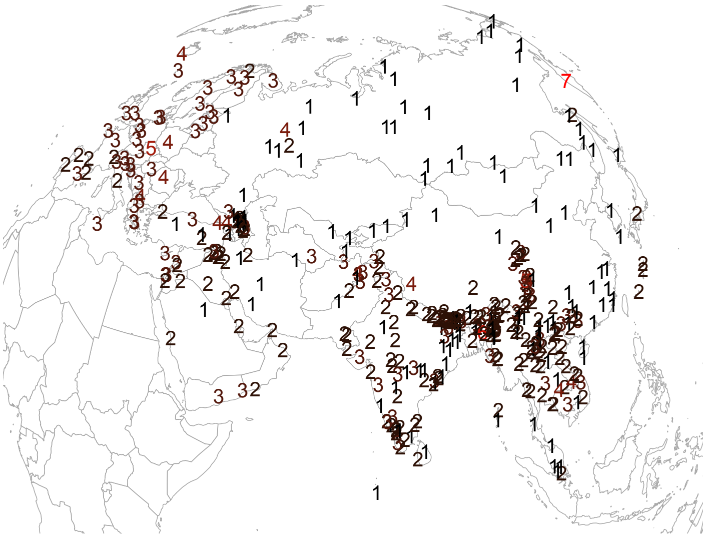

Fig. 2 Maximal length of an onset in each lect

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 the minimal number of onset in each lect, which is either one or zero. We see that the minimal onset length of one, or the obligatory onset, is mostly present in the Mainland Southeast Asian linguistic area (Enfield, 2018; Vittrant and Watkins, 2019; Sidwell and Jenny, 2021, cf) 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. [ʔ] may or may not be a phoneme in such lects.

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]).

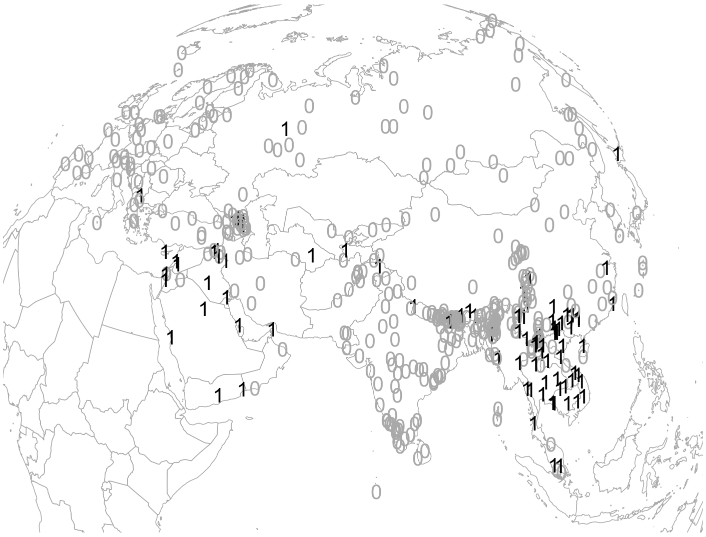

Fig. 3 Minimal length of an onset in each lect

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, 57), Cantonese diphthongs are not analyzed as vocalic sequences within a nucleus but rather 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 offglide is a coda itself.

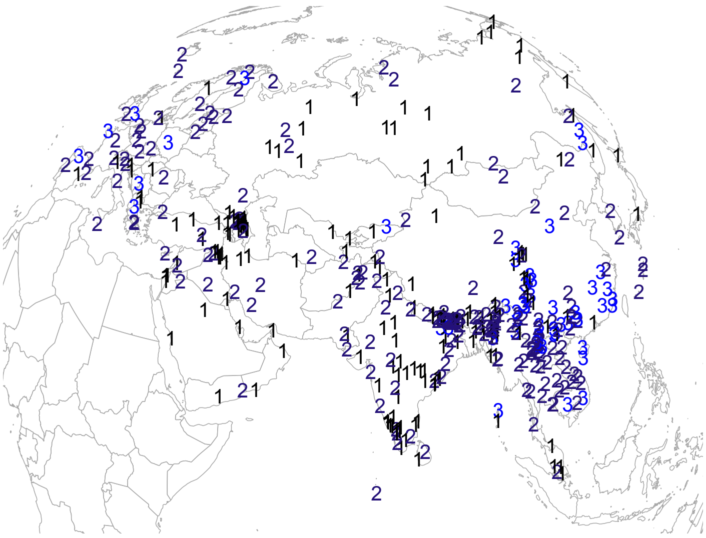

Fig. 4 Maximal length of a nucleus in each lect

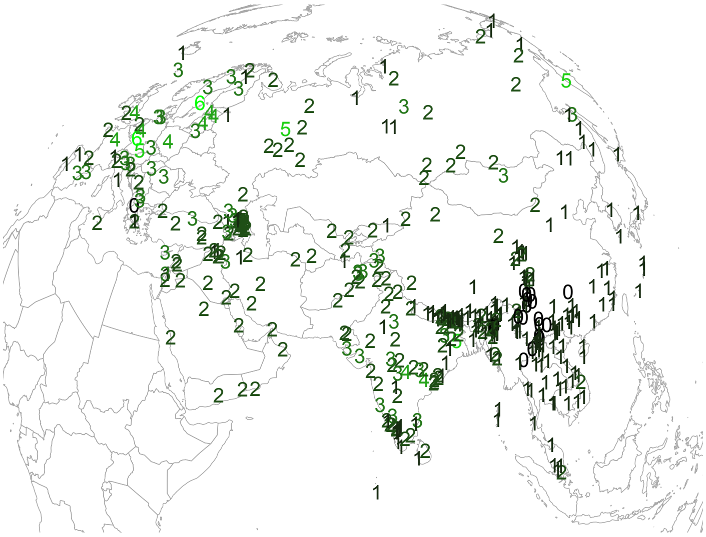

Fig. 5 Maximal length of a coda in each lect

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 are coda-less, not allowing any coda at all. In sum, we observe a general correlation between onset length and coda length in the Eurasian macroarea, which both tend to increase westwards.

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 have 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 1,516km distance. This distance threshold leaves no sample lect without any neighbor. We then created a weight matrix and assign the value of 1 to each neighboring lect pairs and the value of 0 to each non-neighboring lect pairs. Based on this weight matrix, we performed the Moran's I test (Table 4) to test the spatial autocorrelation, which confirms that both onset length and coda length are areally clustered ($p < 0.001$). 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. The results (Table 5) show that both onset and coda lengths are correlated with longitude. This confirms our visual observation that the maximal onset length and the maximal coda length grow as one goes westwards in Eurasia.

Category	Moran I statistic	Expectation	Variance	p
Onset	0.0459892	-0.0023202	0.0000345	< 0.001
Coda	0.3413282	-0.0023202	0.0000345	< 0.001

Table 4 Moran's I

Category	Intercept	Coefficient	SE	p
Onset	2.414	-0.008	0.001	< 0.01
Coda	2.775	-0.012	0.001	< 0.001

Table 5 Spatial regression of longitude and onset/coda length

Other than geographical coordinates, it is worthwhile to compare the maximal length of onset/nucleus/coda based on language families. 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.

6.2 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 its nucleus position, as they allow consonants to form the nucleus. Consonants that form the nucleus are known as the *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

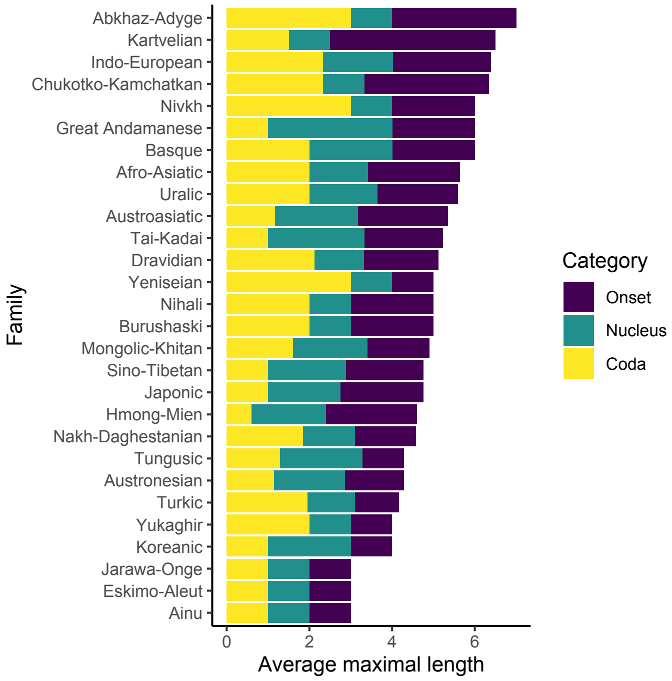

Fig. 6 Average maximal length of onset by family

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.

6.3 Number of singleton codas

In Section 6.1, 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 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

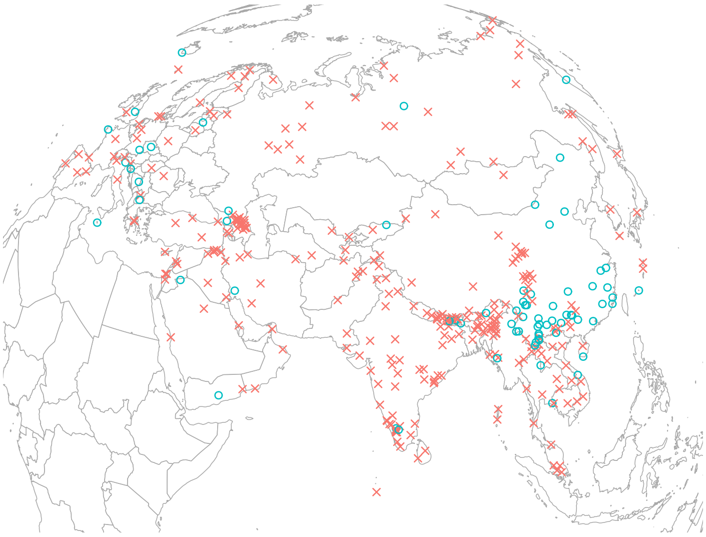

Fig. 7 Syllabic consonants

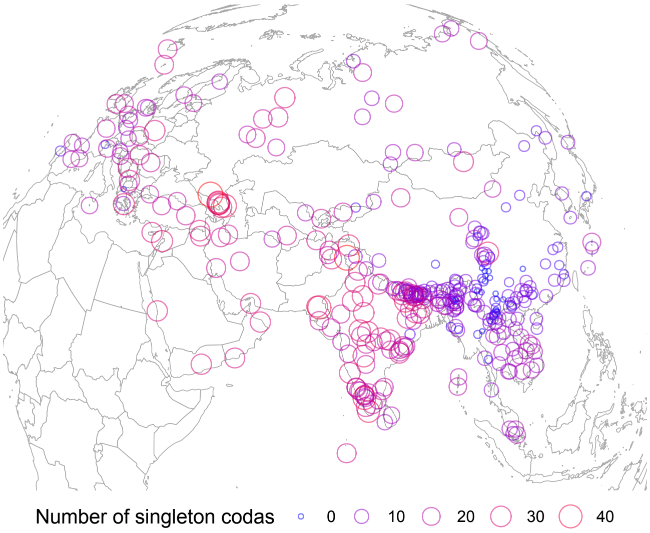

Fig. 8 Number of singleton codas

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 and plosives, and glides are the most common consonants as coda, whereas liquids, fricatives, affricates are less common.

6.4 Number of tones

Maddieson's (Maddieson (2013b)) 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 shows that tonal lects are most heavily present in Sub-Saharan Africa and Mainland Southeast Asia. Complex tone system (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.

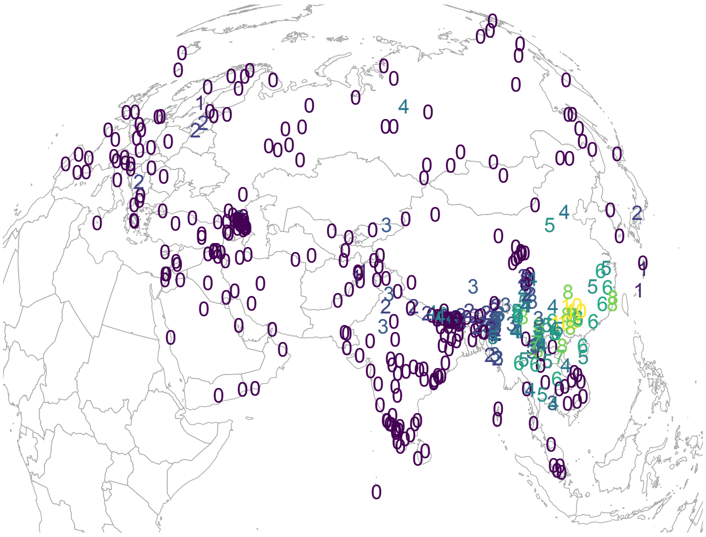

Fig. 9 The number of tones per lect

Figure 9 shows the number of tones per Eurasian lect. The largest number of distinctive tonemes is ten, e.g. in Cao Miao (Wu, 2015), and the lowest number is one, e.g. 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 (2013b). 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, with at most two tones. From this uneven distribution, we can know 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 its tones inherited from Middle Korean in certain varieties (notably the Southeast variety), while the Seoul variety is currently going through Tonogenesis (Kang and 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.

6.5 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 (§6.4), for example, is not identical to the distribution of syllabic consonants (§6.2). It is therefore helpful to shed light on each one of the phonological patterns to understand their diverse areal shapes.

7 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 will be 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. 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 many researches on phonological diversity and universality in Eurasia and beyond.

References

Anderson C, Tresoldi T, Greenhill SJ, et al (2021) Measuring variation in phoneme inventories. https://doi.org/10.21203/rs.3.rs-891645/v1

Anderson GD, Harrison DK (1999) Tyvan. Lincom

Bauer RS, Benedict PK (1997) Modern Cantonese phonology. Mouton de Gruyter

Berg T (1986) The monophonematic status of diphthongs revisited. Phonetica 43(4):198–205

Bickel B, Nichols J, Zakharko T, et al (2022) The autotyp database. https://doi.org/10.5281/zenodo.6793367

Blasi DE, Wichmann S, Hammarström H, et al (2016) Sound–meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Sciences 113(39):10818–10823

Blasi DE, Moran S, Moisik SR, et al (2019) Human sound systems are shaped by post-neolithic changes in bite configuration. Science 363(6432):eaav3218

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

Dixon RMW, Aikhenvald AY (2003) Word: A typological framework. In: Dixon RMW, Aikhenvald AY (eds) Word: a cross-linguistic typology. Cambridge University Press, p 1–41, https://doi.org/10.1017/CBO9780511486241.002

Doornenbal M (2009) A grammar of bantawa: Grammar, paradigm tables, glossary and texts of a rai language of eastern nepal. PhD thesis, Rijksuniversiteit te Leiden

Dwyer A (2013) Tibetan as a dominant sprachbund language: Its interactions with neighboring languages. In: The third international conference on the Tibetan language. Trace Foundation, p 258–280

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

Enfield NJ (2018) Mainland Southeast Asian Languages: A Concise Typological Introduction. Cambridge University Press

Fleischer J, Schmid S (2006) Zurich german. Journal of the International Phonetic Association 36(2):243–253

Fuchs R (2015) Word-initial glottal stop insertion, hiatus resolution and linking in british english. In: Sixteenth annual conference of the international speech communication association

Goldsmith J (2011) The syllable. In: Goldsmith J, Riggle J, Yu ACL (eds) The handbook of phonological theory, 2nd edn. Wiley, p 164–196, https://doi.org/10.1002/9781444343069.ch6

Gowda KSG (1968) Descriptive analysis of soliga. PhD thesis, Deccan College

Grossman E, Eisen E, Nikolaev D, et al (2020) Segbo: A database of borrowed sounds in the world's languages. In: Proceedings of the 12th language resources and evaluation conference. European Language Resources Association, p 5316–5322

Gut U (2009) Introduction to English phonetics and phonology, vol 1. Peter Lang GmbH

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

Hammarström H, Forkel R, Haspelmath M, et al (2021) Glottolog 4.4. Max Planck Institute for Evolutionary Anthropology, https://doi.org/10.5281/zenodo.4761960

van der Hulst H, Ritter NA (1999) Theories of the syllable. In: van der Hulst H, Ritter NA (eds) The syllable: views and facts. De Gruyter Mouton, p 13–52, https://doi.org/10.1515/9783110806793.13

Iwasaki S (2013) Japanese, revised edn. John Benjamins Publishing Company

Janhunen J (2006) Sinitic and non-sinitic phonology in the languages of amdo qinghai. In: Anderl C, Halvor E (eds) Studies in Chinese language and culture: Festschrift in honour of Christoph Harbsmeier on the occasion of his 60th birthday. Hermes Academic Publishing, p 261–268

Jenny M, Hnin Tun SS (2016) Burmese: A comprehensive grammar. Routledge

Kahn D (1976) Syllable-based generalizations in english phonology. PhD thesis, Massachusetts Institute of Technology

Kang Y, Han S (2013) Tonogenesis in early contemporary seoul korean: a longitudinal case study. Lingua 134:62–74

Lee 이 J� (2021) Kwuke umwunlon kanguy 국어 음운론 강의 [A course in Korean phonology]. Jipmundang 집문당

Li X, Li J, Luo Y (2014) A grammar of Zoulei (Southwest China). Peter Lang

List JM, Forkel R, Greenhill SJ, et al (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 I (2009) Patterns of sounds. Cambridge University Press

Maddieson I (2013a) Syllable structure. In: Dryer MS, Haspelmath M (eds) The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, URL https://wals.info/chapter/12

Maddieson I (2013b) Tone. In: Dryer MS, Haspelmath M (eds) The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, URL https://wals.info/chapter/13

Maddieson I, Flavier S, Marsico E, et al (2013) Lapsyd: Lyon-albuquerque phonological systems database. In: Interspeech 2013. International Speech Communication Association (ISCA), https://doi.org/10.21437/interspeech.2013-660

Malaia EA, Wilbur RB (2020) Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. Wiley Interdisciplinary Reviews: Cognitive Science 11(1):e1518

Mielke J (2008) The emergence of distinctive features. Oxford University Press

Moran S, McCloy D (2019) PHOIBLE 2.0. Max Planck Institute for the Science of Human History, URL https://phoible.org/

Moran S, Grossman E, Verkerk A (2021) Investigating diachronic trends in phonological inventories using bdproto. Language Resources and Evaluation 55(1):79–103

Mortensen DR, Littell P, Bharadwaj A, et al (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, pp 3475–3484

Nikolaev D (2018) The database of eurasian phonological inventories: A research tool for distributional phonological typology. Linguistics Vanguard 4(1)

Nikolaev D (2019) Areal dependency of consonant inventories. Language Dynamics and Change 9(1):104–126

Pike KL (1947) On the phonemic status of english diphthongs. Language 23(2):151–159

Riad T (2013) The Phonology of Swedish. Oxford University Press

Rzymski C, Tresoldi T, Greenhill SJ, 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 R, Bickel B, Hildebrandt KA (2010) The prosodic word is not universal, but emergent. Journal of Linguistics 46(3):657–709

Sidwell P, Jenny M (2021) The languages and linguistics of Mainland Southeast Asia: a comprehensive guide. De Gruyter Mouton, https://doi.org/10.1515/9783110558142

Skirgård H, Haynie HJ, Blasi DE, 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 S (2000) The Ainu language, 1st edn. Sanseido

van der Hulst H (2017) Phonological typology. In: Aikhenvald AY, Dixon RMW (eds) The Cambridge Handbook of Linguistic Typology. Cambridge University Press, p 39–77, https://doi.org/10.1017/9781316135716.002

Vittrant A, Watkins J (eds) (2019) The Mainland Southeast Asia linguistic area. De Gruyter Mouton, https://doi.org/10.1515/9783110401981

Wu M (2015) A grammar of sanjiang kam. PhD thesis, University of Hong Kong

Xu D (2017) The Tangwang language: An interdisciplinary case study in Northwest China. Springer

Zakaria M (2018) A grammar of hyow. PhD thesis, Nanyang Technological University

Zhou C (2020) Case markers and language contact in the gansu-qinghai linguistic area. Asian Languages and Linguistics 1(1):168–203