PHONOLOGICAL AREAS IN EURASIA

IAN JOO

PhD

The Hong Kong Polytechnic University

2023

The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies

Phonological areas in Eurasia

Ian Joo

A thesis submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy

August 2023

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Joo, Ian (Name of student)

Dedication

To grandmother

할
머
니
께

Abstract

This thesis investigates the phonological areas of Eurasia. A phonological area is a geographical area where different lects have converged into similar phonological patterns. In order to compute the distribution of phonological areas in Eurasia, I have built Phonotacticon 1.0, a database consisting of basic phonotactic information of more than 500 Eurasian lects. It includes the segmental phonemic inventory, tonemes, and onset/nucleus/coda sequences of each sample lect. I employ this database to measure the phonological distance between Eurasian lects and clustering them to detect areal patterns within Eurasia. The phonological convergence patterns generated thereby largely overlap with the previously hypothesized linguistic areas, such as Europe, Northeast Asia, Qinghai-Gansu, or Mainland Southeast Asia. This dissertation thus presents a novel method to measure the similarity between two phonological structures and use that method to confirm the linguistic areas previously argued for.

Acknowledgements

I would like to thank first and foremost my chief supervisor, Dr. Yu-Yin Hsu, for her kind support and inspiring discussions.

My sincere gratitude also goes to my co-supervisor, Prof. Chor Shing David Li.

I have stayed at Uppsala University, Sweden, during the first three months of 2022, as an exchange student. I thank Prof. Harald Hammarström for hosting me and providing me rich resources and comments for building Phonotacticon.

I would also like to thank my colleagues at the Hong Kong Polytechnic University – whose names I cannot exhaustively list here – for the wonderful time I have enjoyed with them during my doctoral years in Hong Kong.

Special thanks to fellow typologists and phonologists at the LingTyp mailing list and Twitter, most of whom I have never met, but whose comments have provided me considerable feedback for my doctoral research.

Lastly, my thesis, my academic career, and I as a person would not have become complete if not for the immense love and support from my family: My mother, my father, my brother, and especially my grandmother, to whom this thesis is dedicated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

When different human societies meet, the members of each society are exposed to each other's different lect[1], via trade, migration, education, intermarriage, language shift, or other forms of cultural exchange. This phenomenon is called *language contact*. Language contact usually leads the lects in contact to develop similar linguistic patterns, such as shared vocabulary or syntactic isomorphy. This process is known as *linguistic convergence*.

Throughout human history, language contact and linguistic convergence have mostly occurred between geographically close peoples. Due to the physical limits of human transport and communication, especially in pre-modern times, the majority of human interaction has occurred between human groups within geographical vicinity. What naturally follows, then, is that lects that are used in geographically adjacent regions have come into more contact and developed more convergence than lects that are geographically far apart.

Due to the geographical bias of language contact, a geographically adjacent group of lects often develop a significant level of convergence and form a geographical space characterized by a certain set of shared linguistic features. Such space is a *linguistic area*. Well-known linguistic areas include Europe (Haspelmath 2001), South Asia (Masica 2005), and Ethiopia (Bisang 2006). A *phonological area* is a subset of linguistic area, limited to the domain of phonology. Linguistic convergence may occur in one domain but not in another: Two lects may develop a significant degree of similarity in their phonology but not in their morphosyntax or lexico-semantics. A phonological area is a type of linguistic area, which does not imply the existence of morphosyntactic or lexico-semantic areas in the same geographical space.

In this thesis, I compute the distribution of phonological areas in Eurasia. By Eurasia, I refer to the macroarea as defined by Hammarström and Donohue (2014), which is largely the

[1] In this thesis, I use the term *lect* to refer to any level of linguistic variety, commonly referred to as a *dialect* or a *language*. This is because the distinction between a dialect and a language is inherently sociocultural and not language-internal, and thus not relevant for the current research. I only use the term *language* when I refer to the general concept of the human language, such as in *language contact*.

same as (but not identical to) the Eurasian continent. In order to detect phonological areas, it is necessary to first compute *phonological distance*, the degree of difference between two phonological structures. For example, I must quantify how English phonology is different from French phonology compared to Mandarin phonology. Based on the distance between each pair of Eurasian lects, I am able to cluster the Eurasian lects into groups of phonologically similar lects and map those clusters onto a geographical plot to verify whether phonological clusters correspond to geographical clusters.

To achieve this goal, I build *Phonotacticon*, a database that consists of phonological information of spoken lects (not including signed lects). Phonotacticon includes the following information of 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).

For my doctoral project, I have compiled the Eurasian part of Phonotacticon, consisting of ca. 500 lects. This Eurasian part of the database, or *Phonotacticon 1.0*, is available at `https://github.com/ianjoo/Phonotacticon`. In this dissertation, I present the building of Phonotacticon 1.0, use this database to compare the phonological distance between Eurasian lects, and use the distances to detect phonological areas in Eurasia.

1.2 Research goals

The goal of this thesis is twofold:

- **To build a phonological database that contains the basic phonotactic information of Eurasian spoken lects.** How can we build a database containing the different phonotactic rules of hundreds of lects in a cross-linguistically consistent manner?
- **To use this database to calculate the phonological distance between Eurasian lects and thereby cluster them to test if phonological clusters form geographical clusters.** Can we quantify the phonological distance between the sample lects and thereby measure one distance against another? If we cluster the sample lects based on their phonological distance, do the clusters show geographical patterns?

In order to achieve these two goals, the remaining part of this thesis takes the following steps:

- Chapter 2 reviews previous literature relevant to this thesis.
- Chapter 3 shows how I built Phonotacticon 1.0 (first goal).
- Chapter 4 shows some descriptive visualizations based on Phonotacticon, such as the distributions of onset/nucleus/coda length and the number of tones.
- Chapter 5 uses Phonotacticon 1.0 to measure the phonological distance between Eurasian lects and cluster them (second goal).
- Chapter 6 concludes the thesis.

Chapter 2

Literature review

2.1 Introduction

In this chapter, I will review some of the previous studies relevant to my goal of using a phonological database to measure cross-linguistic phonological distance in order to detect phonological areas (geographical areas of phonological convergence) in Eurasia. Section 2.2 introduces the concept of *phonological convergence*. Section 2.3 discusses the notion of *linguistic area*, focusing on those in Eurasia. Section 2.4 summarizes previously built *phonological databases*. Section 2.5 reviews previous methodologies of quantifying *phonological distance*. Section 2.6 concludes by summarizing how these previous studies are relevant to this thesis.

2.2 Phonological convergence

Phonological convergence is the phonological domain of *linguistic convergence*, the assimilation between two or more lects via language contact. In other words, phonological convergence is the assimilation between two phonological **structures** (which must be distinguished phonological **forms**, cf. §2.5) due to language contact. When two lects come into contact, they often develop a phonological pattern that resembles that of the other. Such phonological pattern can be individual phonemes, phonotactic restrictions, syllable structures, or phonological rules.

There is reason to analyze phonological convergence independently from other types of convergence, such as morphosyntactic or lexical convergence. Previous literature suggests that linguistic convergence may be domain-specific: that is, convergence in one domain, such as syntax, does not imply convergence in another, such as phonology. Meakins and Pensalfini (2021) show that two Australian lects, Jingulu and Mudburra, share a great deal of mutually borrowed vocabulary but retain each of their distinct grammar. François (2011) demonstrates how northern Vanuatu lects (all belonging to the Oceanic branch of the Austronesian family) have phonologically and lexically diverged but show a great degree of syntactic isomorphism. Donohue (2013, p. 223) takes Basque and Dravidian lects as examples where the dominant Indo-European lects

have affected their phonology and Khoi-San as an example of receiving morphosyntactic influence from the Niger-Congo superstratum. Thus, phonological convergence between two lects may not imply their convergence in other domains, or vice versa.

One of the mechanisms of phonological convergence is lexical borrowing. When a lect imports a considerable amount of loanwords from another lect containing a sound pattern not present in the recipient lect, then that sound may develop into part of the recipient lect's phonology. An example is the /ʒ/ in German, attested mostly in French loanwords like *Genie* /ʒɛni/ 'Genius' or *Garage* /gaʁaʒ/ 'garage' (Wiese 2000, p. 12). According to Wiese (2000, p. 12), as there is no tendency to assimilate /ʒ/ into any one of the other native German phonemes, it should be considered an integral part of German phonology.

A set of externally adopted sound patterns may form part of a *phonological stratum* of a lect. A good example is Japanese, whose lexical strata consist of native Japonic, Sino-Japanese, non-Sinitic loans, and ideophones (Itô and Mester 1999). The four strata obey different phonotactic rules. Native Japonic does not allow word-initial /r/, whereas it is permitted in all other three strata (e.g. Sino-Japanese *ryō* 寮 [rjoː] 'dormitory'). [ɸV] sequences other than [ɸɯ] are only attested in non-Sinitic loans (e.g. English loan *fan* ファン [ɸaɴ] '(someone's) fan'). Thus, although loan sounds can be a true part of the recipient lect's phonology, it often forms a distinct layer within the phonological system.

Loanwords are not the only origin of external import of phonological patterns, however. Yurok, an Algic lect spoken in Northwestern California, has the ejective consonants /ʧʼ kʼ kʷʼ pʼ tʼ/, which are not found in its sister lect Wiyok (Blevins 2002). Blevins (2017, p. 96) argue that although /tʼ ʧʼ/ are found in loanwords, /pʼ kʼ/ are almost non-existent in loanwords and thus better analyzed as internal innovation due to areal pressure from neighboring lects with ejectives.

Blevins' (2017) *stone soup theory* is a theory of such phonological convergence that is internally processed but externally motivated. The European fable of the stone soup is a story of a visitor tricking their hosts into making a "stone soup". First, the visitor pretends to be able to cook delicious soup with stone as the only ingredient. As they are cooking, they suggest to their hosts that a little bit of certain (edible) ingredients would make the soup even better. After convincing the hosts to share their ingredients multiple times, the visitor succeeds in cooking their "stone soup", whose taste in fact originates from the ingredients shared by the hosts and not from the stone.

This fable, argues Blevins (2017), is analogous to the internally processed but externally motivated phonological convergence. As the stone in the fable is what attracts the hosts to create the soup, the areal feature in the neighboring lects of a lect is what attracts that lect to develop that feature within itself. The stone is the external motivation and not the actual ingredients. Likewise, even if a sound change can happen fully internally within a lect, it can still be externally motivated by the lects it has contact with.

An important aspect of phonological convergence (and other domains of linguistic conver-

gence) is the continuity of convergence. Externally imported sound patterns do not simply enter the recipient lect at once, but rather blend into it gradually, first at its periphery, then slowly towards its core. While /ʒ/ can be recognized as a phoneme assimilated to the core of German phonology, nasal vowels attested in French loanwords (e.g. *Restaurant* [ʀɛstoːʀɑ̃]) still remain at the periphery of German phonology, as not all German speakers pronounce such loanwords with nasal vowels (Wiese 2000, p. 12). In this sense, we can say that a phonological system of a lect is not a discrete category but rather a prototypical category, some members (sound patterns) of it being closer to the prototype (core), while some are further away.

This continuity of convergence applies not only to sound patterns that are adopted but also to the speaker population who adopt them. That is, different speakers within a group of a recipient lect may accept the imported sound pattern at different levels. Chirkova et al.'s (2018) study on the phonological convergence of Ersu (Sino-Tibetan) towards Southwestern Mandarin, such as the simplification of complex onsets, shows how the convergence patterns can manifest at different levels depending on the speaker's sociocultural background, such as their occupation or education level. In other words, phonological convergence happens not abruptly but gradually, **sound by sound** and **speaker by speaker**.

Lastly, we should not forget that language contact not only causes linguistic convergence but also linguistic divergence (Kühl and Braunmüller 2014; Evans 2019). Naturally, contact may induce phonological divergence as well. In Temiar (Austroasiatic), some Malay loanwords go through phonological processes not attested elsewhere in Temiar phonology, such as final denasalization (Malay /kəbun/ > Temiar /kəbut/ 'orchard'), the sole purpose being signaling their foreign origin, as the Temiar culture wants to distinguish them from native lexicon (Benjamin 1976). Although the present thesis will focus on phonological convergence, given that language contact can also cause phonological divergence, any absence of phonological convergence shown in the following chapters should not be immediately interpreted as absence of contact.

2.3 Linguistic area

In this section, we turn to one model of linguistic convergence, the *linguistic area*. A linguistic area is a geographical area home to multiple languages that share a number of linguistic features due to historical contact and not genealogical relationship. In other words, it is a geographical group of linguistic convergence.

Section 2.3.1 briefly introduces the concept of linguistic area. In the remaining subsections, I will introduce some of the major linguistic areas in Eurasia that have been proposed and argued for by previous works. The following chapters of this thesis will investigate whether the analysis based on Phonotacticon support the previously claimed linguistic areas.

2.3.1 What is a linguistic area?

The concept of linguistic area, or *Sprachbund* (GER 'language union'), was first formally defined by Trubetzkoy (1928) as a group of lects sharing a high number of morphosyntactic, phonological, and lexical similarities but no regular sound correspondence in their morphological elements or basic vocabulary (and thus cannot be traced back to a common proto-lect). Trubetzkoy's definition was above all meant to clearly distinguish a linguistic area from a language family, consisting of a group of lects sharing a common ancestor. As an example, he cites Bulgarian as belonging to the Slavic family (which in turn is a branch of the Indo-European family) but belonging to the Balkan linguistic area along with Greek, Albanian, and Romanian.

Two implications in Trubetzkoy's brief definition must be highlighted. First, Trubetzkoy does not include genealogical unrelatedness as a criterion of a linguistic area. Most of the lects spoken in the Balkan Peninsula, like Bulgarian, belong to the Indo-European family. From his definition, it seems that although linguistic area and language family must be conceptually distinguished, members of a linguistic area do not have to belong to different language families.

Second, Trubetzkoy does not mention geographical proximity as a criterion, although many studies following his (Thomason 2000, e.g.) define a linguistic area as a geographical area. It is noteworthy that he named the concept "language **union**", unlike its English calque "linguistic **area**", suggesting that he did not see this concept as a geographical space but rather a relationship between lects. It is indeed possible for two lects spoken very far away from each other to come into contact and develop similarity: Malay, for example, have developed some degree of lexical and phonological similarity to Arabic (namely the adoptation of Arabic xenophones, such as /f/ or /x/), even though these two lects are spoken in distant regions, due to the religious influence of Islam in the Malay Peninsula. This type of long-distance contact is the exception rather than the rule, however. As most linguistic contacts happen in geographical vicinity, it does not create much problem to view a linguistic area as a geographical area in most cases.

Thomason's (2000) definition of a linguistic area is perhaps more pertinent to how the term is used in linguistics today: "A linguistic area is a geographical region containing a group of three or more languages that share some structural features as a result of contact rather than as a result of accident or inheritance from a common ancestor" (p. 311). Her definition captures the main criteria of a linguistic area: (i) geographical region; (ii) three or more lects; and (iii) similarity due to contact. Note that none of these three terms were included in Trubetzkoy's original definition. Very technically, according to his definition, a sprachbund could also consist of two lects spoken in distant regions that share similarities by chance.

The first criterion, geographical region, was not highlighted by Thomason herself, but remains an important aspect of the contemporary definition of a linguistic area. Humans live within geographical boundaries, be they mountains, rivers, oceans, jungles, deserts, or geopolitical borders. Conceptualizing linguistic area as a geographical area hosting different lects rather than viewing it as a set of lects per se emphasizes the spatial nature of linguistic contact

and turns the agenda of linguistic area research into the discovering areal patterns on human-inhabited space rather than solely investigating similarities between different lects.

The second criterion, three or more lects, adds importance to the multidimensionality of the contact that constitutes a linguistic area. Thomason (2000, p. 312) suggests "perhaps the major reason for considering two-language contacts separately from [linguistic areas] is that in the great majority of the cases the source of a shared feature is easier to determine when only two languages are involved". But it is not due to practicality of research alone that a bilateral contact must be distinguished from a multidimensional zone of contact. Conceptually, a space arises only when there are more than three dots connected. Two dots can only form a line between the two. A contact between two lects can only be understood in terms of bilateral relationship between the two, unlike the multiangular connection between three or more lects, which forms a complex dimension of contact dynamics and may be conceptualized as an area. Thus, when we say that three or more lects are required for a linguistic area, three is not just an arbitrary number but represents a fundamental distinction between bilateral and multilateral contacts.

The third criterion, similarity due to contact, rules out any similarities that arose due to common inheritance or simple chance. Shared traits due to descending from a common proto-lect must be distinguished from shared traits due to contact, even though these two are often hard to distinguish when the lects in contact belong to the same family. For example, the southern Sinitic lects share many traits commonly inherited from Middle Chinese, such as tones, as well as areal features not inherited from Middle Chinese, such as the merger of initial /n-/ and /l-/ (Huang 2007). Moreover, two lects can be similar in some aspects plainly due to chance: Ainu and Malay, for example, are highly similar phonologically, as I will show in Chapter 5. It would of course be absurd to assume that Ainu and Malay are somehow distantly related as a family or that the Ainu and the Malay peoples have had some kind of unknown contact in the distant past. Their phonological similarity, thus, is best explained as accidental. As difficult as distinguishing inheritance from areality or accidence can be, the contact-induced origin of the shared features is "the whole point of the concept" of a linguistic area (Thomason 2000, p. 312).

Thomason (2000) also includes "structural features" as one of the criteria of a linguistic area. By "structural", she excludes shared vocabulary as a valid areal feature of a linguistic area. The reason for this is that if we count loanwords as a possible shared feature of an area, "then the entire world would be one linguistic area, thanks to such widely shared words as *email, hamburger, democracy, pizza, Coca Cola,* and *television*" (p. 312).

While the cultural loanwords like those Thomason (2000) gave as examples are indeed transmitted quite easily and may not form a valid criterion of areality, the basic vocabulary of a lect, such as body part terms, are harder to change due to contact. Thus, common lexico-semantic patterns in basic words that arose via contact can be rightfully regarded as shared features of a linguistic area. Brown's (2013) survey of lects colexifying (using the same lexeme for) HAND and FINGER show that this colexification is concentrated in Australia and North America. Schapper et al. (2016) shed light on the unusual colexification between FIRE and FIREWOOD in Australia

and New Guinea and suggest it to be an areal feature. There seems to be no reason to exclude the lexicon from the criteria of a linguistic area, although the distinction between basic vocabulary and cultural and technical vocabulary must be made.

What is unclear, as Thomason (2000, p. 313) points out, is how many shared features are necessary to constitute a linguistic area. There is no consensus on the absolute number required, nor there should be. This is partly because each feature weighs differently based on their typological rarity. The feature of having /f/ does not weigh the same as having click consonants, as /f/ is far more common typologically than click consonants. Thus, we could say that the one feature of having click consonants can outweigh multiple ordinary features like having /f/, allowing a coda, or being tonal. It is thus up to individual researchers to decide how many shared features are enough to argue for the linguistic area they hypothesize.

Panov (2020) makes an important distinction between *unique areal features* versus *non-unique areal features*. That is, a feature does not have to be unique to a geographical area in order to be a characteristic of that area. For example, tonality is an important feature of the Mainland Southeast Asian linguistic area (§2.3.2.3). It is not an exclusive feature of that area, however, as at least a third of the world's lects are tonal (Maddieson 2013b). Thus, tone is a non-unique areal feature of the Mainland Southeast Asia.

What make a feature areal, then, is not necessarily its exclusivity, but rather its absence in the regions surrounding it. In this sense, an areal feature may be described as a dot on a paper. For ink to form a dot on a paper, there does not need to be only one dot on the whole sheet. There needs to be, however, at least some space surrounding the dot absent of ink. Otherwise, there would be nothing perceivable as a dot. As put by Chrikba (2008), "it is not necessary that a certain linguistic be a unique property of this particular zone not found beyond its boundaries", but it is necessary "that this trait, even if not unique in itself, is specific enough to make a meaningful contrast with languages outside this area" (p. 27).

In sum, a linguistic area can be defined as a geographical area of multiple lects sharing a certain amount of contact-induced convergence patterns not shared by their neighbors surrounding the area.

2.3.2 Linguistic areas in Eurasia

It is impossible to tell how many linguistic areas exist in Eurasia. One reason is that some of the proposed linguistic areas are disputed, such as the Caucasian linguistic area (Chirikba 2008, cf.). Another reason is that linguistic areas are multi-layered: A large linguistic area can nest a smaller linguistic area, such as the Balkan linguistic area within the European linguistic area. Thus, the number of linguistic areas in Eurasia depends on what linguistic theories to accept and how fine the areal resolution should be.

In this section, I will present some of the larger linguistic areas in Eurasia that have been proposed or adopted by multiple researchers.

2.3.2.1 Northeast Asia

Northeast Asia is the northeasternmost corner of the Eurasian continent consisting of northeast China, Mongolia, Siberia, Russian Far East, Korea, and Japan. A few researchers have suggested the Northeast Asia to be a linguistic area, without much consensus on what the main common features are or where the geographical boundaries lie.

Hölzl (2018, p. 8) defines Northeast Asia as the part of Eurasia that is "north of the Yellow River and east of the Yenisei". He views Siberia and Qinghai-Gansu (cf. §2.3.2.2) as sub-areas of Northeast Asia, further questioning whether these two regions form independent areas themselves at all. While a few researchers has argued Siberia to be a linguistic area (G. D. S. Anderson 2006; Georg 2008; Vajda 2008), whether it is at the same layer as Northeast Asia is unclear and most researchers include at least some portion of Siberia in their definition of Northeast Asia. Whether Qinghai-Gansu is a subset of Northeast Asia is even less clear, but the data from the present thesis shows that it is phonologically distinct from other lects of Northeast Asia and may form a phonological area at the same level as Northeast Asia (§5).

Whitman (2016), based on linguistic features retrieved from the World Atlas of Language Structures (Dryer and Haspelmath 2013), conducted multiple correspondence analysis on 201 sample lects. Based on his phylogenetic clustering, Kolyma Yukaghir, Evenki, Khalkha Mongolian, and Turkish form one cluster. Another cluster is formed by Burmese, Japanese, Korean, Ainu, and Nivkh. These two clusters, along with Kannada and Meithei, together form one branch. Except for Turkish, Burmese, Kannada, and Meithei, all these lects are spoken in Northeast Asia. Three other Siberian lects, Ket, Nenets, and Chukchi, were not included in the Northeast Asian cluster. As Nenets is spoken in Western Siberia and Ket along the Yenisei basin, this concurs with Hölzl's (2018) definition of Yenisei being the western limit of Northeast Asia. Chukchi is spoken in the northeasternmost edge of Northeast Asia, suggesting that the Northeast Asia as a linguistic area does not reach as far northeast as Chukotka. Moreover, Mandarin is clustered quite distantly from other Northeast Asian lects, despite being geographically spoken in Northeast Asia, and is clustered together with Mainland Southeast Asia lects (cf. §2.3.2.3) – Khmer, Thai, and Vietnamese – and Yoruba (Atlantic-Congo). This concurs with the present dissertation's results (Chapter 5) showing that Mandarin is phonologically similar to Mainland Southeast Asian lects than to other Northeast Asian lects, suggesting that the northern limit of Mainland Southeast Asia may reach as far north as Beijing, which in other words would form the southern limit of Northeast Asia.

Szeto and Yurayong (2021), based on thirty linguistic features, show that northern Sinitic lects are closer to "Altaic" lects (as a typological group consisting of Turkic, Mongolic, and Tungusic families) than southern Sinitic lects are, which are closer to Mainland Southeast Asian lects. The Altaic-like features of northern Sinitic include the retroflex fricative initial (e.g. /ʂ-/ in Mandarin) and distinction between plain negative marker and existential negative marker (e.g. plain negative *bù* 不 and existential negative *méi* 沒 in Mandarin). There's no doubt that within

the Sinitic spectrum, northern Sinitic lects are closer to the non-Sinitic lects of Northeast Asia than southern Sinitic lects are. It is important to note, however, that the thirty features used as the parameter by Szeto and Yurayong (2021) are mostly features that are specifically selected to highlight the north-south contrast of Sinitic. In other words, while Szeto and Yurayong (2021) show that northern Sinitic is more Altaic and less Mainland Southeast Asian **when compared to southern Sinitic**, it does not follow that northern Sinitic is closer to Altaic **than it is to Mainland Southeast Asia**. If the Altaic-ness of southern Sintic was, say, 10% and its Mainland Southeast Asian-ness 90%, the Altaic-ness of northern Sinitic could be 30% and its Mainland Southeast Asian-ness 70%, which would make northern Sinitic more Altaic than southern Sinitic is but still more Mainland Southeast Asian when compared to its Altaic-ness.

Yurayong and Szeto (2020), based on forty linguistic features, show that while many Northeast Asian lects, including Turkic, Mongolic, Tungusic, Chukotko-Kamchatkan, and Nivkh, do form a typological cluster, Japonic, Koreanic, and Ainu are typologically distinct from them. Sinitic, including northern Sinitic, is distinct from both Northeast Asia and Japonic/Koreanic/Ainu. Based on their results, it is possible that the Northeast Asia as a linguistic area does not reach the Korean peninsula and the Japanese archipelago. Overall, the boundaries of Northeast Asia as a linguistic area remain difficult to define.

2.3.2.2 Qinghai-Gansu

The Bodic, Turkic, Sinitic, and Mongolic languages spoken in Qinghai and Gansu province of western China form together the *Qinghai-Gansu linguistic area*, also known as the *Amdo Sprachbund*. Although Amdo Tibetan and Northwest Mandarin serve as the two lingua francas (Dwyer 2013, p. 264), contact-based influences betwen all the four families are attested. While the geographical mass of Qinghai-Gansu is far smaller than other areas discussed in this chapter, it displays a distinct mixture of linguistic features that is hard to define as either Northeast Asia or Mainland Southeast Asia.

Xu (2017, Ch. 1) lists five features common to Qinghai-Gansu:

(i) Verb-final word order

(ii) Case marking

(iii) Terminative suffix *thala*

(iv) Inanimate plural marking

(v) Converbs

Dwyer (2013, p. 66) lists four features that are present in most lects:

(i) CV(N) syllable structure

(ii) ONE as the postpositive indefinite article

(iii) Tense-aspect as verbal suffixes

(iv) Bodic vocabulary for animal husbandry, hunting, and Tibetan Buddhism

Particularly notable is the case-marking of Sinitic (C. Zhou 2020), which is rarely attested elsewhere. Examples in (1) show the case-marking in Linxia Chinese (Peyraube 2017, slightly modified):

(1) a. 我 這個 人哈 認不的
Wo zheige ren-ha renbude
1SG this.CL person-ACC not.know
'I don't know this person.' (Accusative)

b. 北京-ta 回來 了
Beijing-ta huilai le
Beijing-ABL return PRF
'They are back from Beijing.' (Ablative)

c. 他 晌午-tala 睡 了
Ta shangwu-tala shui le
3SG midday-ALL sleep PRF
'He slept until midday.' (Allative)

d. 我 筆兩個 寫 去
Wo bi-liangge xie qu
1SG pen-INS write go
'I am writing with a pen.' (Instrumental)

In the phonological domain, Janhunen (2006) observes that the lects of this area have either Bodic or Sinitic phonology. Both types of phonology are syllable-based with strong coda restrictions, the main difference being that the Bodic type allows complex onsets. Turkic and Mongolic influence on phonology, according to Janhunen (2006, p. 263), has mostly disappeared over time. Thus, in the Qinghai-Gansu linguistic area, we can say that the dominant morphosyntactic models are Turkic and Mongolic, whereas the dominant phonological models are Bodic and Sinitic.

An example of a non-Bodic lect adopting Bodic phonology is Wutun (Sinitic), which has lost tones and developed voiced obstruents due to the influence from Amdo Tibetan (Sandman 2016). It also allows the velar nasal as an onset (e.g. [ŋu] 'I'), which is a character of Northwest Mandarin not found in Beijing Mandarin (Sandman 2016, p. 31). According to Chen (1988), Wutun had complex onsets as well. It allowed /h ʕ n ŋ m/ as possible preinitials (e.g. /hʤɑ/ 'grass', /ŋgon/ 'temple'), much like Amdo Tibetan (Ebihara 2019). Sandman (2016, p. 35) reports that these preinitials are now lost, however. Nevertheless, this suggests that Wutun's Bodic character was even stronger before.

An example of Sinitic phonology of a non-Sinitic lect is the phonology of Mangghuer (Mongolic; Slater 2003). Although Mangghuer phonology has both Sinitic and Bodic characteristics, Slater (2003) views Sinitic as the primary driving force of Mangghuer's phonological innovations. Sinitic characters of Mangghuer include the retroflex consonants /ʂ ʈ͡ʂ ʈ͡ʂʰ ɻ/, which are typical to northern Sinitic. The simplicity of ts syllable structure also resembles Sinitic, the maximal syllable template being CGVC, where the coda is restricted to sonorants. Dwyer (2008) also reports the ongoing tonogenesis in Mangghuer.

Shared lexicon is also a characteristic of the Qinghai-Gansu area. Eastern Yugur (Mongolic) and Western Yugur (Turkic), both spoken by the Yugur ethnic group, share a large set of common vocabulary borrowed from each other and also from Bodic and Sinitic (Nugteren and Roos 1996). Baonan (Mongolic) spoken in Qinghai has approximately half of its vocabulary borrowed from Tibetan, whereas Baonan spoken in Gansu has much less Tibetic vocabulary (ca. 10%) but more than 40% of its vocabulary borrowed from Chinese, despite the two varieties being mutually intelligible (Wu 2003).

As mentioned in Section 2.3.2.1, some researchers (e.g. Hölzl 2018) include Qinghai-Gansu in the greater area of Northeast Asia as its subarea. But the results shown in Chapter 5 shows that the phonological characters of Qinghai-Gansu are not typically Northeast Asian nor typically Mainland Southeast Asian, suggesting that it should be distinguished from the Northeast Asia as a whole, at least in the domain of phonology.

2.3.2.3 Mainland Southeast Asia

The Sino-Tibetan, Austroasiatic, Austronesian, Tai-Kadai, and Hmong-Mien lects spoken in Indochinese peninsula and Southwestern China form the *Mainland Southeast Asian linguistic area* (Enfield 2018; Vittrant and Watkins 2019; Sidwell and Jenny 2021b). Some of the major features shared by the lects of this area include highly complex tones, monosyllabic or sesquisyllabic lexicon, analytic morphology, and SVO word order. Comrie (2007), based on 21 features selected from the World Atlas of Language Structures (Dryer and Haspelmath 2013), observes that these features point to common patterns in Mainland Southeast Asian languages, whence he concludes that Mainland Southeast Asia is a coherent linguistic area.

The exact boundaries of the Mainland Southeast Asia, if there are any, are a matter of debate. Sidwell and Jenny (2021a) exclude Malay from Mainland Southeast Asia, for it "retains much of its inherited [Austronesian] typology" (p. 3) rather than having converged into the Mainland Southeast Asian features. If Malay, spoken at the Southern end of the peninsula, does not belong to the Mainland Southeast Asian linguistic area, then the Malay peninsula may be the southern limit of the linguistic area.

The northern limit of the Mainland Southeast Asia is less clear. Previous works (de Sousa 2015; Szeto and Yurayong 2021) agree that Far Southern Sinitic lects, spoken in Guangxi, Guangdong, and Hainan, resemble the core members of Mainland Southeast Asia. But it would be hasty to draw the northern boundary of Mainland Southeast Asia based on Sinitic data alone,

as not only Sinitic lects are spoken in southern China. In Yunnan, for example, the local Sinitic lect is a variety of Mandarin, due to relatively recent immigration from northern China. Nevertheless, Yunnan is undoubtedly a part of the core Mainland Southeast Asia, given that the non-Sinitic lects spoken in Yunnan, such as Nuosu (Sino-Tibetan; Gerner 2013, cf.), are genealogically and typologically close to the lects spoken in the Laos or northern Vietnam. Whether the South-Central Chinese provinces, such as Guizhou, Sichuan, or Hunan, belong to this linguistic area is ambiguous. In other words, South-Central China could be Mainland Southeast Asia's (fuzzy) northern boundary.

2.3.2.4 South Asia

South Asia, largely equivalent to the Indian subcontinent, is a linguistic area dominated by Indo-Aryan (⊆ Indo-European) lects in the north and Dravidian lects in the south, while also home to many Sino-Tibetan and Mundaic (⊆ Austroasiatic) minority lects and the lect isolates Nihali and Burushaski.

One of the most prominent areal features of South Asia is the wide distribution of retroflex consonants. PHOIBLE 2.0 (Moran and McCloy 2019) shows that retroflex plosives and sonorants almost exclusively occur in South Asia within Eurasia. Retroflex fricatives and affricates, on the other hand, are not widely distributed throughout South Asia but common in China. /ʂ/ is an exception, as it is common in both regions.

The Indo-Aryan retroflex consonants, attested in the earliest records of Sanskrit, may be an areal influence from the Dravidian substratum, as they are not found elsewhere in Indo-European (Emeneau 1956, p. 7). Although the emergence of retroflexion in Sanskrit can be traced back to internal changes in Indo-Aryan (Arsenault 2012, §2.2.3), even internal changes can result from external influence (Blevins 2017), meaning that its internality does not rule out its areality. Retroflex consonants are also attested in the Mundaic (Arsenault 2012, §2.2.4) and Sino-Tibetan (Arsenault 2012, §2.2.5) lects spoken in South Asia, also likely to be areal influence from Indo-Aryan and Dravidian.

Emeneau (1956) argues that the numeral classifier is an areal feature of South Asia, e.g. Telugu *enimidi mandi manuṣulu* ఎనిమిది మంది మనుషులు <eight-CLF-people> 'eight people'. Moral (1997), however, limit this areal feature to Northeast India rather than South Asia as a whole, claiming that the use of numeral classifiers is limited in other parts of South Asia, especially so as one gets further away from Northeast India. According to Moral (1997), Sino-Tibetan is the source of this feature, as it is common throughtout the Sino-Tibetan family as well as other lects of East and Southeast Asia.

Masica (2005) highlights several morphosyntactic features characteristic of this area, namely:

(i) Head-finality (SOV word order, postpositions, Adj-N/Gen-N/Dem-N/Num-N)

(ii) Morphological causatives (often including double causatives)

(iii) (Heavy usage of) converbs

(iv) Explicator compound verbs, e.g. Hindi *le jānā* ले जाना 'to take away, lit. to take and go'

(v) Dative-subject construction to express possession (rather than using HAVE-like verbs)

Abbi (2018) illustrates *echo formation* as an areal feature of South Asia. Echo formation is a type of reduplication where the base is partially modified in the reduplicant, e.g. Hindi *cāy* चाय 'tea' > *cāy vāy* चाय वाय 'tea and related items'; Tamil *puli* புலி 'tiger' > *puli kili* புலி கிலி 'tiger and others'. Abbi (2018) notes that the echo formations of different South Asian lects are not only morphologically similar but also semantically so. She posits the following semantic functions of South Asian echo formation:

(i) Generality and plurality

- Hindi *pen* पेन 'pen' > *pen ven* पेन वेन 'writing instruments'

(ii) Superordinate structure

- Bangani (Indo-Aryan) *šakun* 'meat' > *šakun-šhukun* 'non-vegetarian, meat related'

(iii) Pejoration

- Hindi *likhnā* लिखना 'to write' > *likhnā vikhnā* लिखना विखना 'to scribble'

(iv) Intensification

- Punjabi *siddhā* सिॱधा 'straight' > *siddhā suddhā* सिॱधा सुॱधा 'absolutely straight'

(v) Sets and types

- Punjabi *nīlā* ठीला 'blue' > *nīlā šīlā* ठीला श़ीला 'blue types'

(vi) Non-specific reference

- Hindi *kənāḍā* कनाडा 'Canada' > *kənāḍā vənāḍā* कनाडा वनाडा 'Canada or some Western country'

A lexical areal feature of South Asia is the richness of ideophones. An ideophone is "[a] member of an open lexical class of marked words that depic sensory imagery" (Dingemanse 2019, p. 16). It is also known as expressives, mimetics, or onomatopoeia (although onomatopoeia is a subset of ideophones, as onomatopoeics depict only sounds). Examples of South Asian ideophones are Maithili *gam gam* 'aroma', Hindi *cam cam* चम चम 'glittering', and Punjabi *las las* ल़म ल़म 'sticky' (Abbi 2018, pp. 12–13). Given the scarcity of ideophones in Indo-European other

than Indo-Aryan and also the systematic similarity between Indo-Aryan and Dravidian ideophones, Emeneau (1969) concludes that the ideophones in Indo-Aryan must be areal influence from Dravidian, without ruling out that Mundaic could have played a role as well.

In sum, there is ample evidence pointing to South Asia as a linguistic area, in the domain of phonology (retroflex consonants), morphology (echo formation), syntax (head-finality, converbs, and dative-subject construction), and lexico-semantics (ideophones).

2.3.2.5 Europe

Europe is the westernmost region of the Eurasian continent delimited from the rest of Eurasia by the Ural mountains, the Caspian Sea, and the Black Sea. Linguistically, it is dominated by various branches of the Indo-European family (Germanic, Italic, Balto-Slavic, Celtic, Hellenic, and Albanian), along with a number of Uralic lects, the Afro-Asiatic lect Maltese, and the lect isolate Basque. Several researchers have analyzed Europe as a linguistic area whose common features cannot be explained by Indo-European inheritance alone.

It was Whorf (1944) who first coined the term *Standard Average European* to refer to the typical model of a European lect. The focus of Whorf (1944), however, was not to linguistically define Europe per se, but rather to highlight the differences between European lects and Hopi, a Uto-Aztecan lect spoken in Arizona, to argue for claims of linguistic relativity. He claimed that Hopi (unlike European lects) does not express time, which is related to the (alleged) absence of the concept of time in Hopi culture. Whorf's (1944) claims about Hopi and linguistic relativity are not accepted today, as Malotki (1983) demonstrated that Hopi does express time in diverse manners, like all human lects do.

Even though Whorf's (1944) theory was unsuccessful, his concept of "Standard Average European" survived and a number of researchers have tried to define the typical features of an "average" European lect. Haspelmath (1998) lists eleven features of Standard Average European:

(i) Definite and indefinite articles (e.g. English *the/a book*)

(ii) Have-perfect (e.g. English *I have eaten*)

(iii) Participial passive (e.g. English *I am seen*)

(iv) Derivation of anticausative from causative (e.g. French *coucher* 'to put to sleep' > *se coucher* 'to go to sleep')

(v) Nominative experiencers (e.g. English *I like this book*) as opposed to dative experiencers (e.g. Hindi *Mujhe yah kitāb pasand hai* मुझे यह किताब पसंद है 'id., lit. This book is preferred to me)

(vi) Dative external possessors (e.g. French *Je me lave les mains* 'I wash my hands, lit. I wash myself the hands')

(vii) Negative indefinite pronoun + verb to express negation (e.g. English *Nobody knows*)

(viii) Particle comparatives (e.g. English *I'm taller than you*) as opposed to surpass comparatives (e.g. Cantonese *Ngo gou gwo nei* 我高過你 'id., lit. I tall-surpass you')

(ix) A and-B conjunction (e.g. English *spring, summer, fall, and winter*) as opposed to A-and B conjunction (e.g. Korean *pom-kwa yelum-kwa kaul-kwa kyewul* 봄과 여름과 가을과 겨울<spring-and summer-and fall-and winter> 'id.')

(x) Postnominal relative clauses introduced by an inflecting relative pronoun, signaling the head's role (e.g. German *der Mann, den ich kenne* <the man, PRON.MASC.ACC I know> 'the man that I know')

(xi) Verb fronting in polar questions (e.g. German *Weißt du das?* <know you that?> 'Do you know that?')

Haspelmath (1998) argues that these eleven features cannot be common inheritance from Proto-Indo-European, as they were absent in Proto-Indo-European, except for dative external possessors. They are thus more likely to be areal innovations.

Note, however, that none of the eleven features are phonological. Haspelmath (2001, p. 1493) also acknowledges the difficulty of finding phonological features common to Europe, suggesting that large vowel inventories and consonant clusters are possible candidates. In Chapter 4, I will show that there are in fact phonological features of Europe that distinguish it from its surrounding areas in Eurasia: First, many European lects allow three or more segments in the onset, nucleus, and coda position. Second, syllabic consonants are present in many European lects. Both features are nearly absent in parts of Eurasia adjacent to Europe (West Asia and North Asia).

2.3.3 Phonological areas

As mentioned in Section 2.2, linguistic convergence may be domain-specific. Phonological convergence may happen with little or no morphosyntactic convergence, and vice versa. It follows that linguistic areas – the geographical areas of linguistic convergence – may also be domain-specific, i.e. there may be "linguistic areas" consisting of lects that have converged in one domain but not necessarily in another. The scope of this thesis remains at linguistic areas in the domain of phonology. i.e. the phonological area. Phonological areas, of course, may overlap with morphosyntactic or lexico-semantic areas – and I suspect that many of them do – but I limit my analysis to claiming that certain phonological areas exist in Eurasia while remaining agnostic about linguistic areas in other domains. In order to detect the existence of phonological areas, I will use a phonological database that I have built, Phonotacticon 1.0. The following section will review previously existing phonological databases.

2.4 Existing phonological databases

In this section, I will review eight of the most important phonological databases, focusing on those that are currently accessible.

2.4.1 UCLA Phonological Segment Inventory Database (UPSID)

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

2.4.2 The Database of Eurasian Phonological Inventories (EURPhon)

The Database of Eurasian Phonological Inventories or EURPhon (Nikolaev 2018, accessible at `eurphon.info`) describes the phonological inventories of 536 Eurasian lects. For many lects, it also contains some phonotactic information, namely word-initial consonant clusters, word-final consonants, and possible syllabic templates. It is perhaps 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 difference. I will explain how they are different in Section 3.5.

2.4.3 PHOIBLE 2.0

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

2.4.4 PBase

PBase (Mielke 2008, accessible at `pbase.phon.chass.ncsu.edu`), provides the following phonological information of 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 t͡ʃ k ʔ b d d͡ʒ 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 on 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 my knowledge, PBase is the only phonological database to distinguish marginal inventory from core inventory and to provide phonological rules such as allophonic variations. Although the marginality of phonemes in any lect is a continuous feature rather than a categorical one, some phonemes being less marginal than others, it is nevertheless highly useful to have a binary distinction of marginal v. core inventories, as the two inventories often behave differently in phonotactic terms. Phonological rules can also provide fruitful information on cross-linguistic phonological patterns, as many phonological rules are shared by different lects, such as final devoicing.

But the highly uneven distribution of the phonological information of different lects makes it less suitable for a quantitative cross-linguistic comparison. 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 lects in Phonotacticon lack one or more of these five variables as well.

2.4.5 Lyon-Albuquerque phonological systems database (LAPSyD)

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

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

For example, the information on Ainu is as follows:

- Segmental inventory: /i u ɛ ɔ a p t̪ k t̪ʃ h s m n ɽ w j/
- Diphthongs: None
- Syllable structures: "(C)V(C). /h/ and /tʃ/ do not occur in coda."
- Consonant notes: "Refsing (1986) describes /t/ as "dental", but /n/ as "alveolar"; /s/ is not specified except that it is [ʃ] before /i/. Simeon (1969) and Refsing (1986) agree in referring to <r> as "post-alveolar" and that it is a flap word-medially, but Refsing labels it "a lax voiced plosive (like Japanese /r/) even sometimes approaching [ɖ]". The transcription chosen here is /ɽ/."
- Vowel notes: "Simeon (1969) and Refsing (1986) both suggest that the mid vowels are lower mid. Refsing (1986, p. 68) describes /u, o/ as having "slight rounding of the lips" but Simeon (1969, p. 755) notes specifically an unrounded allophone of /u/ before /j/ but describes both /o, u/ as rounded."
- Comment on stress: "According to Patrie (1982, p. 128) Ainu has a pitch-accent similar to Japanese; minimal pairs e.g. /torí/ "bird" vs /tóri/ "stay over"; /niná/ "to knead" vs /nína/ "to gather firewood". High pitch in Hokkaido corresponds to long vowels in the Karafuto dialect of Sakhalin. Refsing (1986, pp. 73–74) reports that the first syllable is usually accented if it is closed, otherwise the second; hence there are few minimal pairs."
- Comment on tone: "Ainu has a pitch-accent similar to Japanese (see "stress")"
- Location: (43, 143), "Parts of Hokkaido, N. Japan. Formerly spoken in Kurile Is. and Sakhalin peninsula, Russia"

Perhaps one of the greatest values of LAPSyD lies in its qualitative details, especially for suprasegmental aspects like stress, which are better explained qualitatively. The great amount of such detailed explanation in verbal format makes this database highly useful for lect-by-lect consultation.

2.4.6 BDPROTO 1.1

BDPROTO 1.1 (Moran, Grossman, et al. 2021, accessible at `github.com/bdproto`) contains the phonological inventory of 257 ancient and reconstructed lects around the world. It is, to my knowledge, the only phonological database with a non-contemporary lects. As Moran et al. (2021, p. 87) point 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 of each proto-lect, making BDPROTO a useful database to conduct a diachronic analysis on phonological typology.

2.4.7 SegBo

SegBo (Grossman et al. 2020, accessible at `github.com/segbo-db`), is a list of the borrowed segments in more than 500 lects worldwide. According to SegBo, /f/ is the most commonly borrowed segment worldwide. As SegBo also codes the donor lect of each segment, it shows that the following five are the largest 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 around the world, especially the asymmetrical contact between less spoken lects and larger, dominant lects.

One of its limits is that SegBo (as described in Grossman et al. 2020) is not areally balanced, overrepresenting certain regions like Papunesia and eastern Russia. East Asian lects are quite underrepresented in the database, hence the underrepresentation of Mandarin Chinese as a donor lect. But as SegBo is still in its early stage, this problem can be easily overcome by adding more sample lects.

2.4.8 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 me know that it will be available online again in the near future.

Personal communication with Siva Kalyan also informed me that 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)*. Although I do not know whether segment-level information will be available once the database is back online, 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.

2.4.9 Summary of phonological databases

Table 2.1 summarizes the eight databases I have reviewed in this section.

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, each lect bearing different values of a set of phonological parameters, and not a form-based database containing possible phonological forms a lect

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	>500	World	Borrowed segments	Yes
WPD	Thousands	World	Phonotactics	No

Table 2.1: Summary of the eight databases reviewed

can generate according to its phonotactic rules. This form-based phonotactic database is what Phonotacticon aims to be. It will contain the basic phonological profiles of lects worldwide, now containing around 500 Eurasian lects.

While Phonotacticon 1.0 may serve various purposes, the primary use of it in this thesis is to measure the phonological distance between Eurasian lects. The following section will introduce the concept of phonological distance and review previous methodologies of quantifying phonological distance.

2.5 Interstructural phonological distance

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

The second meaning is the distance between the phonological structures of two lects. How close is English phonology to Turkish phonology, in terms of phonological inventory, phonotactic constraints, or segmental frequency? And is the distance closer than the distance between English phonology and Japanese phonology? In contrast to the cross-sequential phonological distance, I call this type of phonological distance *interstructural phonological distance*.

Measuring intersequential phonological distance and measuring interstructural phonological distances may share certain processes. The distance between two phonemes, such as the distance between [m] and [p], is relevant to both types of phonological distances: Intersequentially, it is required for measuring how close /mæn/ is to /pæn/; interstructurally, it is required for measuring the distance between a lect with /m/ but without /p/ in its phonemic inventory and a lect with /p/ but without /m/. But as the two distances are measured in two different dimensions - phonological form vs. phonological structure - they must be clearly distinguished in order to

avoid confusion.

Many works on *dialectometry*, the measuring of distance between dialects, involve measuring the phonological distance between dialects. They are better classified as works on intersequential phonological distance rather than interstructural phonological distance. As an example of a work on dialectometry, Flikeid and Cichoki (1987) measured the distance between Acadian French idiolects based on a number of phonological parameters. The phonological parameters mostly pertain to phonetic variants of certain French phonemes, such as whether Standard French /k/ is pronounced as the affricate [tʃ] or whether /u/ is pronounced as the diphthong [uw]. Because the different phonemes of each dialect are compared to common reference forms (Standard French), measuring such phonological distance is in effect measuring how a given sequence in a French dialect will be pronounced differently in another French dialect. It is thus closer to intersequential phonological distance than to interstructural phonological distance.

As one of the two goals of the thesis is to measure the interstructural phonological distances, the following subsections will provide an overview of how previous works have tried to measure it in different ways. The methodological diversity of the previous literature implies that there is no unified mathematical definition of interstructural phonological distance and it is the task of the individuals researchers to measure the distance in their own way.

2.5.1 Avram (1964)

Avram's (1964) four-page paper sketches a methodology to quantify interstructural phonological distance. His measurement is based on the following parameters of each lect:

- the *efficiency* (FRE *efficacité*), the number of phonemes divided by the number of distinctive features;
- the average *distribution* (FRE *distribution*) of distinctive features, where the distribution of a distinctive feature is the number of distinctive features it can co-occur with;
- the average *output* (FRE *rendement*) of distinctive features, where the output of a distinctive feature is the number of phonemes distinguished by that feature; and
- the average *complexity* (FRE *complexité*) of the phonemes, where the complexity of a phoneme is its number of distinctive features.

Avram uses these parameters to compare four genealogically distinct lects: Sanskrit, English, Mandarin, and Nivkh. Although he doesn't provide a general scale of distance, he briefly comments on how the four lects differ in these parameters, such as observing that that Nivkh and Sanskrit phonemes are on average more complex than Mandarin or English phonemes.

Despite being half a century old, Avram's short paper received close to no attention to this day. Some of his novel ideas, such as the phonemic complexity or the featural distribution, merit to be reconsidered for future research on phonological distance measuring.

2.5.2 Postovalova (1966)

Postovalova (1966) provides a method of measuring the *valence* (RUS *valentnost'* *валентность*) of a lect's phonological feature with another feature. The valence of a feature F_1 with the feature F_2 is calculated as follows:

$$\frac{\text{(Number of phonemes with } F_1 \text{ and } F_2\text{)/(Number of phonemes with } F_1\text{)}}{\text{Number of features - 1}} \tag{2.1}$$

As an example, suppose that English has ten distinctive phonological features. English has three phonemes that are [+nasal] (/m n ŋ/) and one phoneme that is [+nasal, +labial] (/m/). The valence of [+labial] with [+nasal] would be calculated as:

$$\frac{1/3}{10-1} = \frac{1}{27} \tag{2.2}$$

On the other hand, all of the three nasal phonemes of English are also [+sonorant]. The valence of [+sonorant] with [+nasal] would then be calculated as:

$$\frac{3/3}{10-1} = \frac{1}{9} \tag{2.3}$$

In other words, Postovalova's valence measures how likely a given feature co-occurs with another feature and weighs it against the total number of features.

Although Postovalova only uses this method to measure the valence of Russian phonemes, she suggests that it could be used for cross-linguistic comparisons as well (p. 35). Later, Afendras (1970) (§2.5.4) adopts her method to measure the distance between Balkan lects.

2.5.3 Kučera and Monroe (1968)

Kučera and Monroe (1968) employ the concepts of *isomorphy* (the correspondence between similar phonemes of different lects) and *isotopy* (the occurrence of similar phonemes in the same syllabic position) to measure the phonological distance between Russian, Czech, and German.

The measure of isomorphy between a lect's set of phonemes P_1 and another lect's set of phonemes P_2 is as follows:

$$1 - \frac{\text{Number of different features between } P_1 \text{ and } P_2}{\text{Largest number of features of any phoneme in either lect}} \tag{2.4}$$

As an example, Kučera and Monroe measures the isomorphy between Russian /b, bʲ/ and Czech /b/. The difference between the two sets of phonemes is 1, because Russian /b/ and /bʲ/ are distinguished by one feature [±sharp], which is absent in Czech /b/. In both Russian and Czech, the largest number of features to define a phoneme is eight. Thus, the isomorphy between Russian /b, bʲ/ and Czech /b/ is as follows:

$$1 - \frac{1}{8} = 0.875 \tag{2.5}$$

Kučera and Monroe paired each set of phonemes of a given lect to the set of phonemes it had the largest isomorphy with (= the phonologically closest). For example, Russian /t/ was paired with Czech /t/, Russian /b, bʲ/ with Czech /b/, Russian /x/ with Czech /x, h/, and so on.

Based on corpora, the authors also calculated the probability of each set of phonemes' occurrence in a given syllabic position. For example, when comparing Russian and Czech, they measured the probability of Russian /b/ or /bʲ/ occurring in the first position of a triconsonantal onset and the probability of Czech /b/ occurring in the same position.

The authors then calculated the *Isotopy Index* (= phonotactic similarity) between lect L_1 and lect L_2 by the following formula:

$$\sum_{i=1}^{n} \frac{2p_i(L_1) \cdot p_i(L_2) \cdot Isomorphy_i}{p_i(L_1) + p_i(L_2)} \tag{2.6}$$

where $p(L_1)$ is the probability that a given set of phoneme will occur in a given syllabic position in lect L_1, $Isomorphy$ the isomorphy between a given set of phoneme in lect L_1 and the corresponding set of phoneme of lect L_2, and n the number of pairs of isomorphic sets of phonemes multiplied by the number of possible syllabic positions.

Based on this measure, they conclude that the Isotopy Index between Russian and Czech (ca. 0.76) is higher than the that between Russian and German (ca. 0.47) or the that between Czech and German (ca. 0.62). This is the expected result, as Russian and Czech both belong to the same Slavic branch of the Indo-European family, whereas German belongs to the Germanic branch.

Kučera and Monroe may have been the first to consider not only the features of the phonemes but also their positional distribution within a syllable. They argue that phonological distance should be measured based on what they name *quantitative phonotactics* (p. 96).

The methodology adopted in this thesis can also be classified as that of quantitative phonotactics. The limit of Kučera and Monroe's approach, however, is that they only considered **in which position a phoneme occurs within a syllable** and not **which phonemes a phoneme co-occurs with in a given position within a syllable**. In other words, Kučera and Monroe only calculated the probability of /s/ occurring in the first position of a biconsonantal onset and compared it to the probability of /s/ (or a similar phoneme) of another lect occurring in the same syllabic position. But they did not consider whether /s/ in this position occurs in /sk/, /sp/, /sl/, or any other combinations of phonemes, which is an important phonotactic variable. If a lect only allows /sp/ and /sk/ as their /sC/ onsets and another lect only allows /sl/ and /sw/, then these two /s/'s of the two lects cannot be regarded as true equivalents, even though they occur in the same position. Chapter 5 will show how I factored this variable into my methodology.

2.5.4 Afendras (1970)

Afendras (1970) slightly modifies Postovalova's (1966) valence model and uses it to measure the distance between lects spoken in the Balkan peninsula, a well-known linguistic area (Joseph 2020). Afendras then visualizes the distance between Balkan lects via *multidimensional scaling*. Although the limit of Afendras's (1970) study is that it only sampled the lects spoken in the same Balkan area, resulting in a multidimensional scale of only one pivot, Section 5.2 will show that multidimensional scaling of hundreds of Eurasian lects will yield multiple clusters.

2.5.5 Eden (2018)

Eden (2018) presents three methodologies for measuring cross-linguistic phonological distance:

- Hamming Distance based on binary phonological features, such as whether each sample lect allows complex onsets or not;
- Entropy algorithm based on IPA-transcribed corpora, or "the relative predictability of a transcribed passage in one language given knowledge of some other language" (p. 193); and
- Spoken language identification based on audio recordings of non-words by participants of different linguistic backgrounds.

Eden uses each methodology to measure the distance between only a few lects, mostly European. But importantly, she illustrates different angles of measuring cross-linguistic phonological distance, suggesting that there is not one solution to this issue but rather many possible approaches, which can complement each other.

2.5.6 Nikolaev (2019)

Nikolaev (2019) presents a novel way to measure the distance between two phonemic inventories, which he name the *Closest Relative Cumulative Jaccard Dissimilarity*.

First, the *Jaccard dissimilarity* between the two phonemes, p_1 and p_2, is defined as follows:

$$Jaccard(p_1, p_2) = \frac{\text{Number of intersect of features}}{\text{Number of union of features}} \tag{2.7}$$

Then, for a lect's phoneme *p*, we identify the phoneme *p'* that has the lowest Jaccard dissimilarity in the lect in comparison.

Finally, the Closest Relative Cumulative Jaccard Dissimilarity between the two lects in comparison, L_1 and L_2, is calculated as follows:

$$\sum_{p \in L_1} Jaccard(p, p') + \sum_{p \in L_2} Jaccard(p, p') \tag{2.8}$$

The higher the Closest Relative Cumulative Jaccard Dissimilarity between two lects, the wider the gap between their phonemic inventories.

As an example, let L_1 be a lect with the phonemes /p, f, m/ and L_2 a lect with the phonemes /p, m/. Let /p/ be defined by the feature [labial], /f/ by [labial, continuant], and /m/ by [labial, voiced, nasal].

The Jaccard Dissimilarity between /p/ and /p/ and between /m/ and /m/ is 0. On the other hand, the Jaccard Dissimilarity between /p/ and /f/ is 1/2, wheras that between /m/ and /f/ is 1/3. Thus, /f/ is the closest phoneme to /p/.

The Closest Relative Cumulative Jaccard Dissimilarity would be thus the following:

$$\begin{aligned}(Jaccard(\text{/p/}, \text{/p/}) + Jaccard(\text{/f/}, \text{/p/}) + Jaccard(\text{/m/}, \text{/m/}))\\ +(Jaccard(\text{/p/}, \text{/p/}) + Jaccard(\text{/m/}, \text{/m/}))\\ = (0 + 1/3 + 0) + (0 + 0)\\ = 1/3\end{aligned} \tag{2.9}$$

The Closest Relative Cumulative Jaccard Dissimilarity method is similar to the measure I will employ in Chapter 5, although there are some differences in detail, namely that I compare onset/nucleus/coda sequences rather than phonemic inventories and that I also factor in negative and neutral featural values (such as [-voi] or [0strid]).

2.5.7 Macklin-Cordes et al. (2021)

Macklin-Cordes et al. (2021) hypothesize that a lect's phonotactic constraints are historically conservative and argue that phonotactic comparison between lects can be used to detect historical phylogeny. They compare different lects "in terms of which sequences of two segments (*biphones*) they permit and which they do not" (p. 225) to detect phylogenetic signals between them. They conclude that phonotactic similarity can demonstrate historical relationship.

The authors briefly mention, however, that one of the limits of their study is that areality was not considered as a potential factor motivating phonotactic similarity (p. 247). This may be a concerning issue when we consider that phonotactics is highly prone to contact-induced change. As I will show in Section 5, genealogically distinct lects within geographical vicinity may develop similar phonotactic structures.

2.5.8 Harnud and Zhou (2021)

Harnud and Zhou (2021) measured the distance between Mongolian (Mongolic), Ewenki (Tungusic), and Uyghur (Turkic) by comparing their vowel qualities. While vowel phonemes may be categorically defined in terms of articulatory features, their precise acoustic characters, such as their duration or their first and the second formants, are continuous variables that differ consid-

erably from lect to lect and from speaker to speaker. For example, although the Japanese vowel /u/ and the Korean vowel /u/ may be transcribed in the same symbol, this hides the fact that Japanese /u/ is articulatorily much less rounded, with only lip compression and no rounding per se (Okada 1991). Harnud and Zhou (2021) used such continuous characteristics of the vowels of the three lects to measure the distance between them. Their results show that in terms of vowels, Mongolian and Ewenki are closer to each other than to Uyghur.

Strictly speaking, Harnud and Zhou (2021) measure the **phonetic** distance between the three lects, rather than their **phonological** distance, as their methodology is based on continuous phonetic data rather than categorical phonological values. But as phonology is essentially based on phonetics (Ohala 1990), it is reasonable to expect that phonologically close lects will also tend to be phonetically closer. Thus, Harnud and Zhou's (2021) methodology may be used in the future to compare the phonetic distance between lects to their phonological distance.

2.5.9 Summary of previous measures of phonological distance

The methodologies I have reviewed in this section measure the interstructural phonological distance in different ways. The diversity of the methodologies suggests that there is no one correct solution to the problem of quantifying phonological distance, but many possible ways. One of those possible ways that I will take in this this thesis (§5) aims to fill some gaps not covered sufficiently by previous works, namely comparing one multisegmental sequence (for example the English complex onset /spl-/) to another, rather than comparing singleton segments.

2.6 Summary

In this chapter, I have reviewed previous literature on the phenomenon of phonological convergence (§2.2), the concept and examples of linguistic area (§2.3), existing phonological databases (§2.4), and previous measures of phonological distance (§2.5). In the remaining part of the thesis, I will use my phonological database to measure the phonological distance between Eurasian lects in order to detect areal patterns of phonological convergence in Eurasia. By comparing the phonological areal clusters generated from my analysis to the linguistic areas discussed in this chapter –Northeast Asia, Qinghai-Gansu, Mainland Southeast Asia, South Asia, and Europe –I will show that my results largely overlap with these five areas, confirming their existence from the phonological perspective.

Chapter 3

Building the database

3.1 Introduction

This section covers the building process of Phonotacticon 1.0, a cross-linguistic phonotactic database of around 500 Eurasian lects. Section 3.2 explains how I chose the ca. 500 sample lects. Section 3.3 lays out the *profile* of each lect coded in the database. Section 3.4 explains what bibliographical sources were consulted for the lects' profiles. Section 3.5 explains how this database is different from an existing database, EURPhon (Nikolaev 2018). Section 3.6 concludes the chapter and preview how the database will be used for the remaining part of the thesis.

3.2 Lect sampling

The 531 sample lects are the lects listed in Glottolog 4.4 (Hammarström, Forkel, 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).

The macroarea "Eurasia" is similar to, but not identical to, the Eurasian continent, as it does not include most southern Pacific islands typically considered to be part of Eurasia, such as Taiwan or Borneo. This macroarea is defined by Hammarström & 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

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

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 531 sample lects is visualized in Figure 3.1, where each color-shape combination represents a family.

Figure 3.1: 531 Sample lects of Phonotacticon

3.3 Phonological profile

Phonotacticon consists of the following phonotactic profile of each of the 531 Eurasian lects:

- P: Phonemic inventory (segmental)
- T: Tonemes
- O: Onset forms
- N: Nucleus forms
- C: Coda forms

Table 3.1 provides an example of the phonological profile of A'ou (Tai-Kadai; Li et al. 2014).

P	p t k q ʔ pʰ tʰ kʰ qʰ d t͡ʂ t͡ɕ t͡ʂʰ 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
T	55 33 13 31
O	p t k q ʔ pʰ tʰ kʰ qʰ d t͡ʂ t͡ɕ t͡ʂʰ t͡ɕʰ m n ŋ l ɭ f s ɬ ɕ χ h v z ʑ ɣ ʁ w j pl bl vl ml
N	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
C	∅ n ŋ

Table 3.1: Phonological profile of A'ou

How were the five variables chosen? The first two variables, the phonemic inventory and tonemes, undoubtedly form the most basic information of a lect's phonology. There is thus little reason not to include them in the database.

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. Previous research shows that the syllable is a neurological phenomenon processed in the superior temporal gyrus (Yu et al. 2015; Oganian and Chang 2019; Yi et al. 2019). 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.

3.3.1 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 voiceless bilabial plosive, <p> overspecifies this phoneme in terms of aspiration, as English /p/ can be either aspirated (as in *pan* [pʰæn]) 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>.

Some grammars employ the concept of **archiphonemes**, a phoneme that represents multiple phonemes. An example is Japanese /N/, which may occur as [n], [m], [ɴ], or others depending on phonotactic context (Iwasaki 2013). As Japanese has both /n/ and /m/ as phonemes, /N/ can be regarded as an archiphoneme that includes these two phonemes. If an archiphoneme occurs in a certain form in isolation, I have transcribed that form to represent the archiphoneme. Japanese /N/ occurs as /ɴ/ when it does not precede any segment (e.g. *pan* パン [paɴ] 'bread'), so I have transcribed it as /ɴ/. When an archiphoneme does not have any isolated form, then I have treated an archiphoneme as equivalent to the phonemes it includes.

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 phonemes 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, I 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 I 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 I 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, I 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, I generally chose the latter as better reflecting the phonology of a given lect.

When transcribing the phonemes based on a reference grammar, I 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 I 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 /ɛ͡i/ as a diphthong phoneme while not having /ɛ/ or /ɪ/ as monophthong phonemes, I still listed /ɛ/ and /ɪ/ as phonemes instead

of /e͡ɪ/.

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 English diphthongs constitute individual phonemes or are combinations of two vowel phonemes is a matter of debate (Pike 1947) and is thus highly subject to theoretical bias. By listing all diphthongs as combinations of monophthong phonemes, I can render Phonotacticon theoretically more consistent. 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, I have made the following changes to PanPhon:

- The features [hitone] and [hireg] were excluded, since they only pertain to tones and not segments.
- I 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>).

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, I have used the capital letters to describe the permitted sequence without specifying the segments: PL for plosive (P) plus liquid (L).

Table 3.2 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.

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, I 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/, I have interpreted that as a phonemic inventory of /p, t, k, n, a, i, u/.

Symbol	Class	Features	Graphemes
B	Bilabial	[+cons, +lab]	
C	Consonant	[+cons]	
Č	Affricate	[+cons, +delrel, -son]	
F	Fricative	[+cons, +cont, -son]	
G	Glide		<j, w, ɥ, ɯ>
Ł	Lateral	[+cons, +cor, +lat]	
L	Liquid	[+cons, +cont, +cor, +son]	
N	Nasal	[+nas, -syl]	
O	Sonorant	[+cont, +son, -syl]	!<h, ɦ>
P	Plosive	[+cons, -cont, -delrel, -son]	
R	Rhotic	[+cons, +cont, +cor, -lat, +son]	
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 3.2: The underspecified segments

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

3.3.2 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, I 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, I locate the syllable boundary according to the following principles:

- If a cluster occurs word-initially or word-finally, then I 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 onset clusters or coda clusters, then I 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 a onset cluster and a coda 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 I resort to the maximal onset principle (Kahn 1976), favoring complex onsets over complex codas. For example, /lpl/ is interpreted as /l.pl/, instead of /lp.l/.
- For triconsonantal or longer medial clusters, I apply the maximal onset principle within the length of the initial cluster. For example, for a medial cluster /lpml/, I 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, I 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, I interpret the word-initial and word-final clusters as the same as onset and coda clusters.

If a consonant is described as occurring word-initially or as an onset, then I 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, I 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 I detect at least one syllable without an onset, then I judge that that language does not oblige an onset.

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

The two vowel symbols <ɪ> and <ʅ> that frequently appear in grammars written in China are interpreted as syllabic consonants /z̩/ and /ʐ̩/, respectively.

Some grammars (e.g. Gowda 1968) treat vowel nasalization as a suprasegmental phoneme rather than treating nasal vowels as phonemes. For theoretical consistency, I have interpreted all such cases as independent nasal vowel phonemes.

3.3.2.1 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 since [t̚] best phonetically represents /t/ than it does /s/, I have listed /t/ as a possible Korean coda but not /s/.

3.3.3 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 I transcribed the non-pitch information in Phonotacticon: <ˀ> for creaky voice, <C> for checked tones, and <^{h}> for aspiration. For example, Burmese tones are transcribed as L (low), Hˀ (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 I assume that the lect has no tone.

3.3.4 Note

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

3.4 Bibliographical source

The phonological information of each lect is derived from the reference grammar classified as the "Most Extensive Description" of each lect in Glottolog 4.4.

When the Most Extensive Description does not depict the modern, Eurasian variety of that lect, have I chosen an alternative reference grammar as the source of the phonological information of a lect. For example, Glottolog lists *Le français au Burundi* [French in Burundi] by Claude Frey as the Most Extensive Description of French. As the Burundian variety of French may be different from modern European French, I have chosen a different reference grammar for French.

I have tried to gain access to the source for every lect, but in some cases, it was realistically impossible to have access, especially when the Most Extensive Description is an unpublished thesis. If I cannot gain access to the Most Extensive Description, I first sought an alternative reference grammar, and if that failed as well, I left out that sample lect from the database. Thus, the database contains slightly less than 531 sample lects.

Moreover, even when the Most Extensive Description was accessible, when the information given was not complete (for example, lacking the full list of consonant clusters) while another source provided more detailed information, I chose the more detailed source, given that I was aware of and had access to it.

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, I 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/ *and any one of* /l r w j/.

3.5 Difference from EURPhon

Although Phonotacticon is similar to EURPhon (Nikolaev 2018), introduced in Section 2.4.2, the two databases are 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.

3.6 Summary

In this chapter, I have introduced the making of Phonotacticon 1.0, a phonotactic database of the Eurasian macroarea. It is the first database containing the possible onset, nucleus, and coda forms of hundreds of lects.

In Chapter 4, I will introduce some visualizations generated from Phonotacticon and discuss areal patterns observable from them. In Chapter 5, I will use the whole database to calculate the phonological distance between the sample lects.

Chapter 4

Descriptive visualizations

4.1 Introduction

Before analyzing the phonological distance between Eurasian lects in the following chapters, I will first present some descriptive visualizations generated from Phonotacticon 1.0. The purpose is to provide a brief overview of diverse phonological patterns across Eurasia, namely syllable length (§4.2), syllabic consonants (§4.3), number of of singleton codas (§4.4), and the number of tones (§4.5).

4.2 Syllable length

In this section, I will visualize the distribution of *syllable length* in Eurasia. By syllable length I 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 my 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

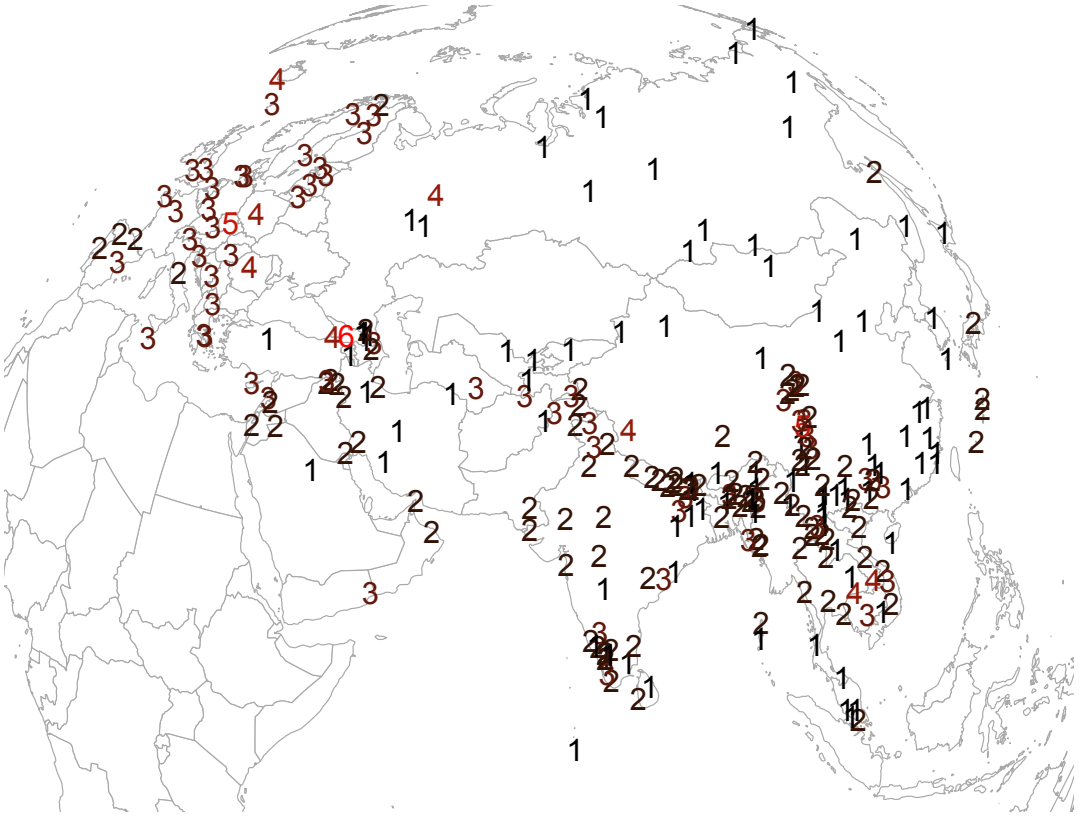


Figure 4.1: Maximal length of an onset in each lect

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 4.1 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 Area (cf. §2.3.2.2); 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.

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 4.2 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 (cf. §2.3.2.3) 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.

Figure 4.3 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

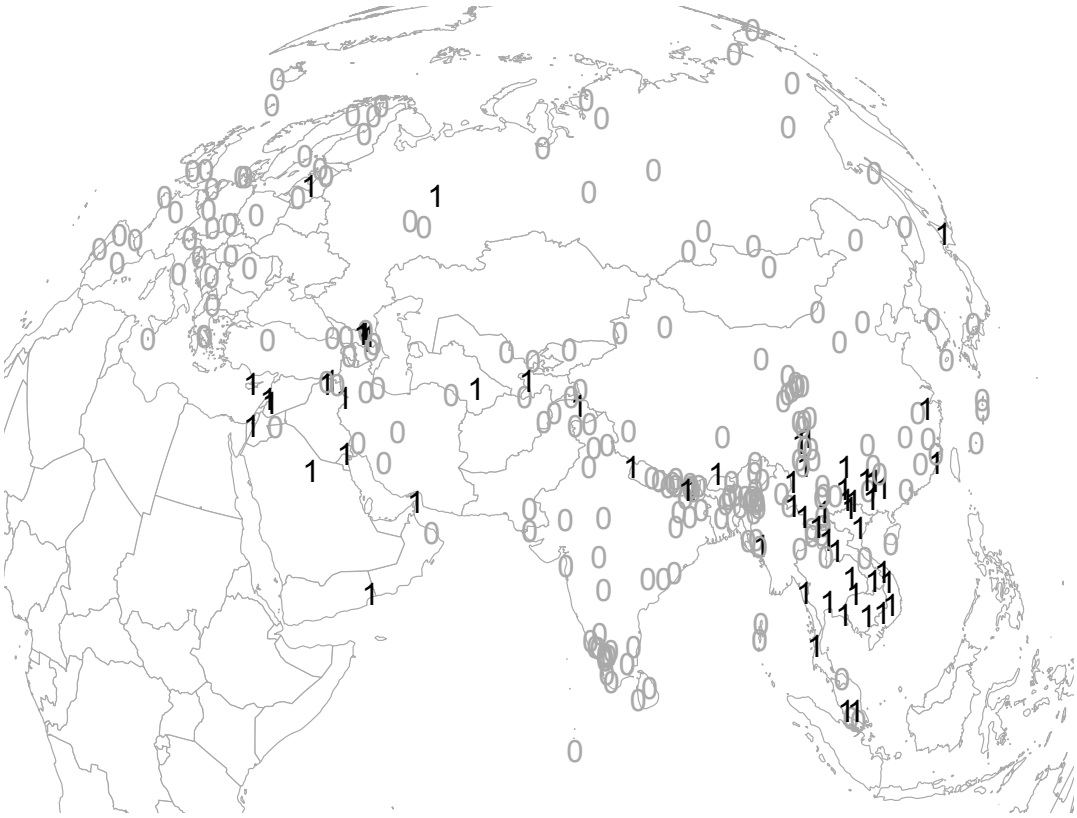


Figure 4.2: Minimal length of an onset in each lect

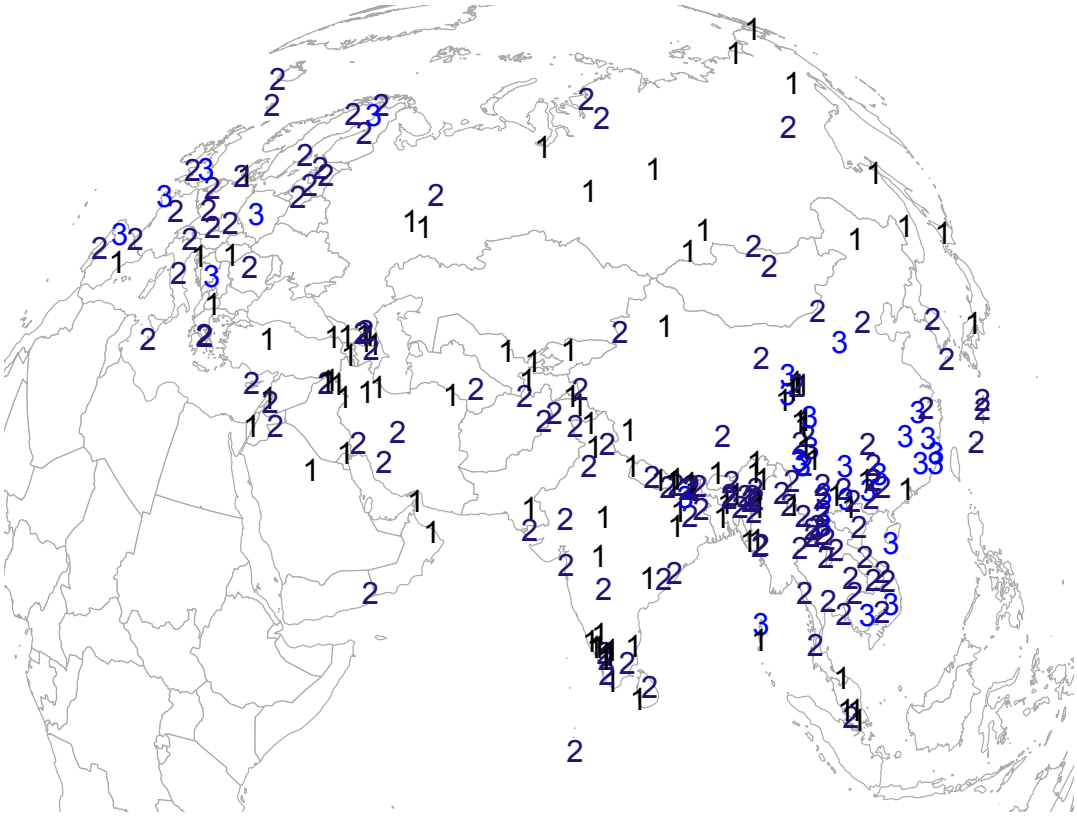


Figure 4.3: Maximal length of a nucleus in each lect

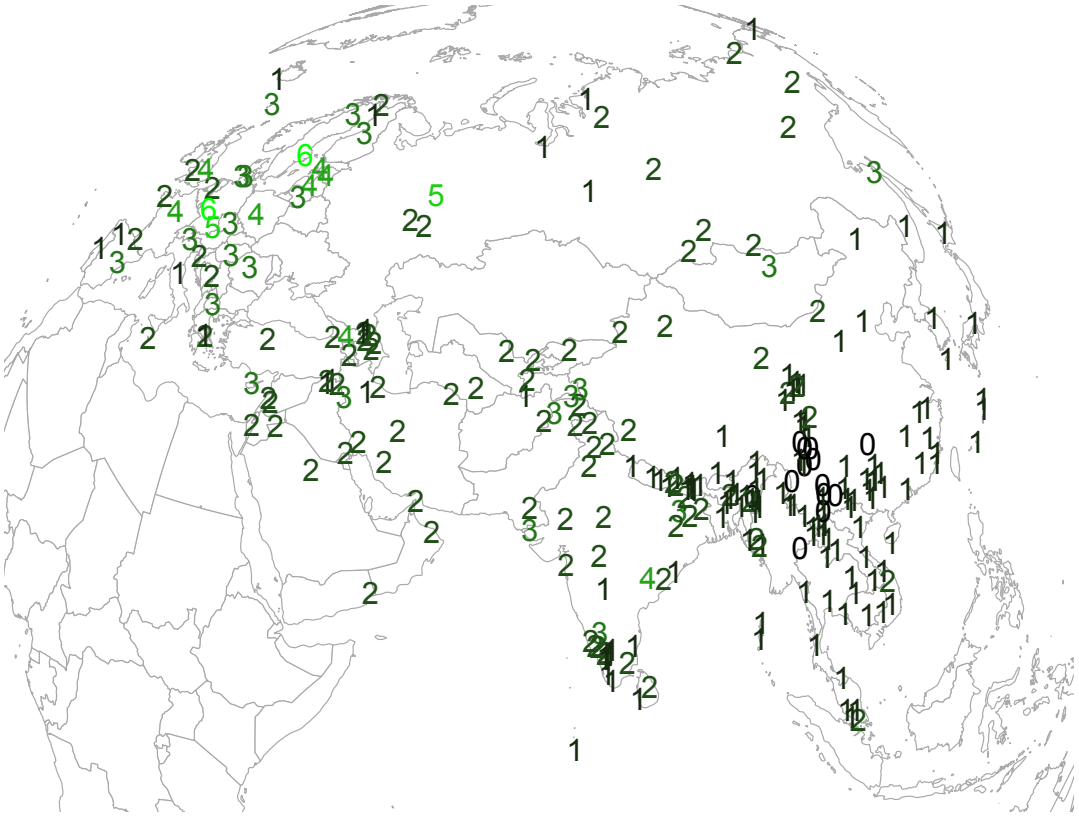


Figure 4.4: Maximal length of a coda in each lect

instance, even though Ainu only permits one vowel in its nucleus slot, glides can appear in the onset or the coda position (/haw/ 'voice'), forming a diphthong beyond the nucleus but within a syllable (Tamura 2000, p. 20).

Figure 4.4 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 4.1: 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.

4.3 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 4.5 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

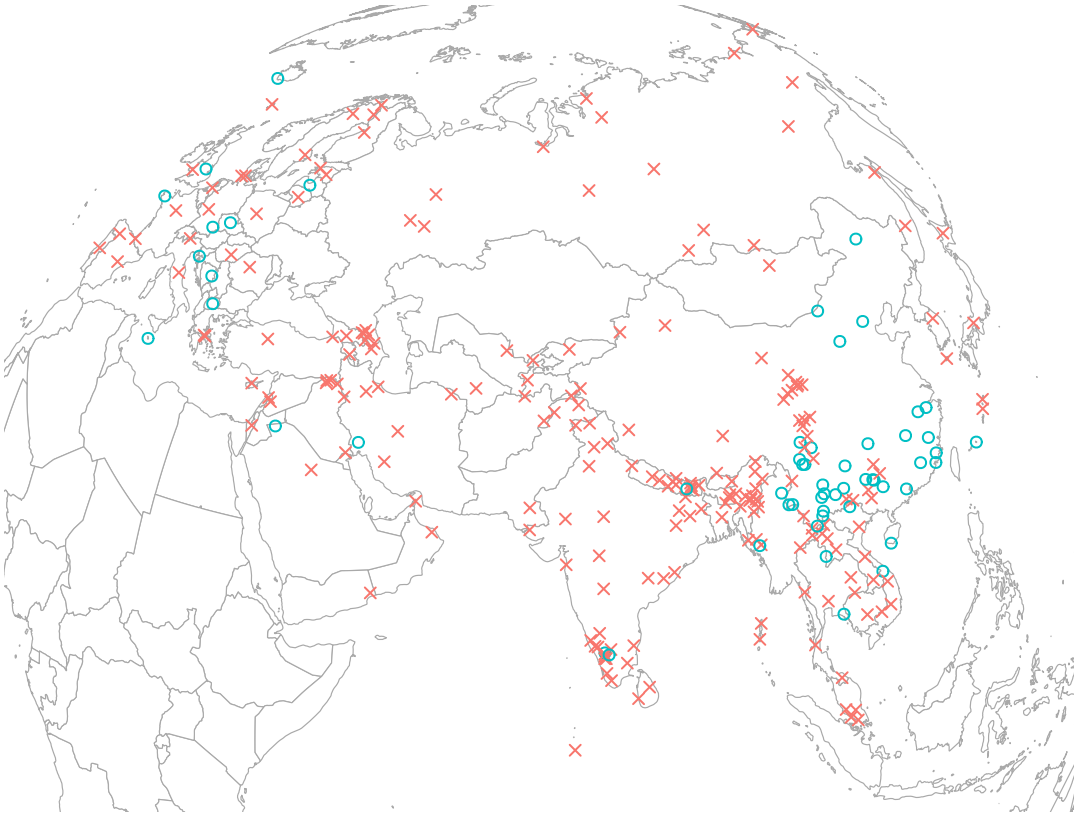


Figure 4.5: Syllabic consonants

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 *m*^{4} 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.tn̩] 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.

4.4 Number of singleton codas

In Section 4.2, 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/

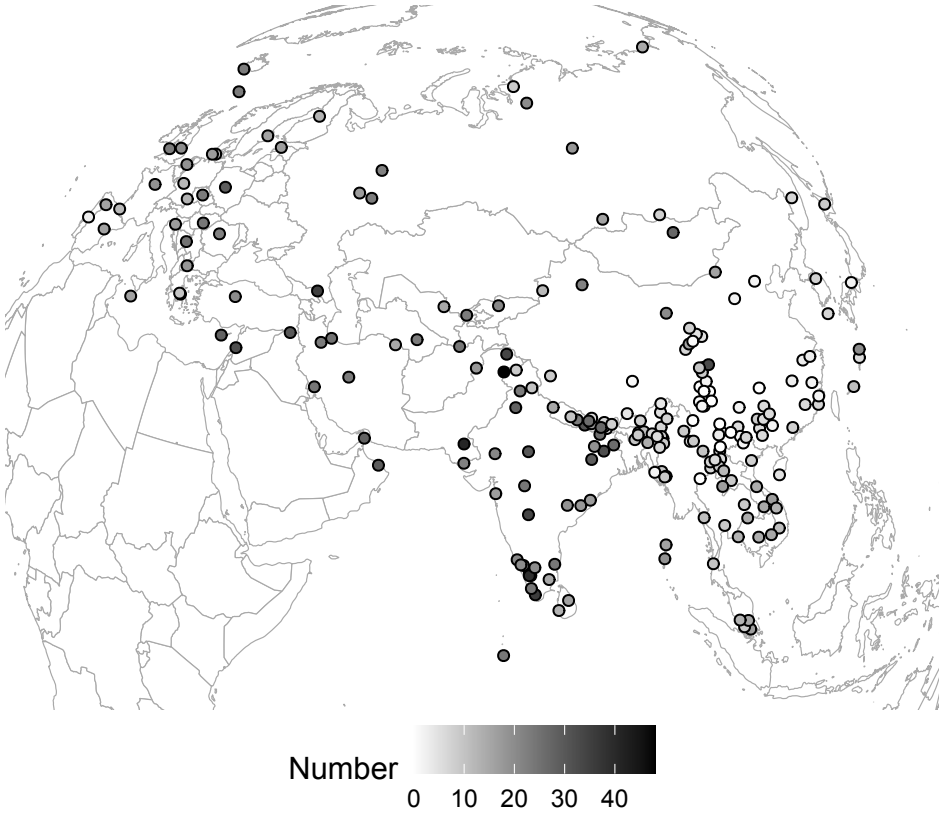


Figure 4.6: Number of singleton codas

as onsets.

Figure 4.6 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 and plosives, and glides are the most common consonants as coda, whereas liquids, fricatives, affricates are less common.

4.5 Number of tones

Maddieson's (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.

Figure 4.7 shows the number of tones per Eurasian lect. It is easily observable that tones are a strongly areal phenomenon, concurring with Maddieson (2013b). Most tonal lects are

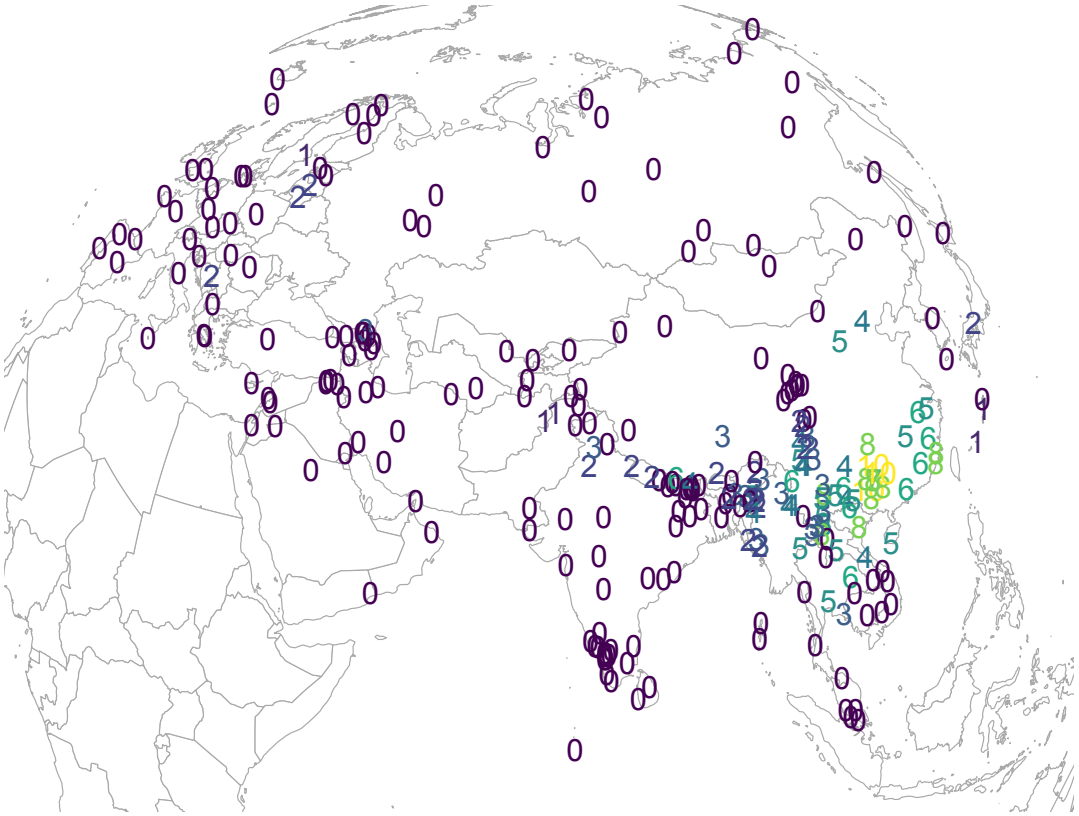


Figure 4.7: The number of tones per lect

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 4.7, 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.

4.6 Summary

In this chapter, we have seen how a number of phonological patterns vary across Eurasia. Crucially, different phonological patterns show different areal distributions: The distribution of tones (§4.5), for example, is not identical to the distribution of syllabic consonants (§4.3). It is therefore helpful to shed light on each one of the phonological patterns to understand their diverse areal shapes.

Phonotacticon can produce many descriptive visualizations as shown in this chapter. The purpose of the database in the context of my thesis is not limited to describing individual phonological features, however. In the next chapter, I will use the database in a comprehensive manner to measure the phonological distance between Eurasian lects and investigate the phonological areas they form.

Chapter 5

Overall phonological distance

5.1 Overview

In the previous chapters, we have seen some interesting visualizations generated from Phonotacticon. They are some of the various ways the database may be used for. My doctoral research's primary objective is, however, to calculate the distance between the phonological profile between lects.

How close is English phonology to French phonology? Is the phonological distance between English and French closer than the phonological distance between English and Mandarin Chinese? Surely, there are many phonological features that English and French have in common but not shared by Mandarin, such as complex onsets, voiced plosives, obstruent codas, and stress-timing. Mandarin also has phonological characters that distinguish it from both English and French, such as lexical tones, retroflex consonants, and syllabic consonants. But how can we quantify these featural differences to compare one distance to another?

The goal of this chapter is to quantify the phonological distance between different Eurasian lects to compare the distance between a pair of lects to the distance between another pair. The ultimate goal is to detect phonologically similar lects within Eurasia to see if they form areal patterns.

Section 5.3 will show the whole process of measuring structural phonological distance by using R (R Core Team 2022) in the format of R Markdown (Xie et al. 2020). Section 5.3 concludes.

5.2 Measuring phonological distance via Phonotacticon

In Section 2.5, we have seen various methods to calculate interstructural phonological distance. In this section, I will present a novel methodology to analyze the phonological distances between Eurasian lects by using Phonotacticon 1.0. While it is similar to Nikolaev (2019) (§2.5.6), the crucial difference lies in that Nikolaev's (2019) distance measure is between phonemic inven-

tories, while the following analysis measures the distance between onset, nucleus, and coda sequences, as well as the numbers of tones.

The following subsections are generated from R Markdown (Xie et al. 2020), which is a human-readable format of the script of R (R Core Team 2022), a programming language designed for data science. In this format, the readers are able to read through the statistical analysis in plain language. The raw code and some of the programming processes were excluded, however, in order to increase readability. Only the crucial parts of the whole programming are shown.

(Continued on the next page)

The data

For computational purposes, lects that are described with underspecified segments (such as "P" for plosives) or sequences with brackets (such as [ptk] [lr] for any sequence with /p/,/t/, or /k/ as the first segment and /l/ or /r/ as the second segment) go through a conversion into segmental information. For example, a lect that has "P" as a possible onset goes through the process of converting "P" into all plosive phonemes it has. A lect that has [ptk] [lr] as a possible onset sequences goes through converting [ptk] [lr] into all the logical possible combinations, i.e. /pl pr tl tr kl kr/.

I excluded sample lects that have sequences including more than two consecutive "C" symbols (representing "consonant"), such as CC or CCC (henceforth CC). This is because such transcriptions would lead to too many possible sequences. For example, one of the English coda sequences is /CCCs/. Obviously, not every logically possible triconsonantal clusters plus /s/ exists in English: It would be absurd to claim that /ssss/ is a possible coda in English. Thus, lects whose sequences include consecutive C's were excluded.

Indeed, sequences using other underrepresentated segments can also generate sequences that do not actually exist in a given lect. For example, Daman-Diu Portuguese (Cardoso 2009) has PL coded as a possible onset sequence, as the source didn't specify whether all the logically possible combinations of the plosives and liquids of its phonemic inventory are present as onset sequences, such as /tl/ or /kl/. But even if these sequences that do not actually exist in a given lect are generated, they are more tolerable than onset sequences coded as CC in other lects, as CC would normally generate far more sequences than PL, which would naturally include far more falsely generated sequences. For example, Amur Nivkh (Gruzdeva 1998) has an onset sequence coded as CC, and as it has 32 consonants, its CC would generate 32 * 32 = 1024 sequences. On the other hand, as Daman-Diu Portuguese has six plosives and two liquids, its PL sequence only generates 6 * 2 = 12 sequences. If a fifth of all the sequences generated from CC in Amur Nivkh or PL in Daman-Diu Portuguese were false, CC would generate approximately 204 false sequences whereas PL would only generate approximately two. As C refers to "any consonant" whereas other underspecified segment symbols (other than V for "vowel") refer to "specific types of consonants (including glides)", it follows that CC would generate far more sequences, and therefore far more falsely generated sequences, than other underspecified sequences.

Moreover, the falsely generated sequences from sequences coded as are phonologically not so distinct from rest of the PL sequences, as they refer to specific subsets of consonants (i.e. plosives and liquids) whereas CC refers to a sequence of any two consonants. For example, even if Daman-Diu Portuguese didn't actually have /tl/ or /kl/, these two falsely generated sequences are not phonologically distant from other plosive-liquid sequences that are attested in the source, such as /tr/ or /kr/, compared to what CC would generate in Daman-Diu Portuguese if it had an onset sequence coded as CC, such as /vf/ or /nʃ/.

Likewise, lects that have bracketed segments with too many members were excluded. For example, one of the possible onsets of Czech is [pbfvmtdszncɟʃʒɲkgxhlr̝j] [pbtdgfvszʃʒxjr̝lmnɲ]. This means any biconsontal sequence whose first member is any one of the 23 segments within the first brackets followed by any one of the 19 segments within the second brackets. As this would generate 437 logically possible sequences, including many onset sequences that do not exist in Czech (such as /pb/, /bt/, or /fm/), it would be overly problematic. All sequences involving ten or more segments within brackets followed by ten or more segments within brackets were excluded.

While a lect having some of its sequences coded as CC or a combination of ten or more bracketed segments is a completely arbitrary variable totally dependent on how its source describes it, given the large number of sample lects, excluding a relatively small number of sample lects that happen to include CC or a sequence of ten or more bracketed segments does not substantially effect the analysis as a whole, whose primary goal is to detect areal patterns across Eurasia rather than focusing on individual lects. Approximately a fourth of the sample lects were excluded for having either CC or a sequence of ten or more bracketed segments.

The following shows the first ten rows and the five columns of PhonoBib, a bibliography of the Eurasian lects containing the bibliographical information as well as the longitude and latitude and language family.

Glottocode	iso	Lect	lon	lat
aoua1234	aou	A'ou	105.8	26.80
abau1245	aau	Abau	141.3	-3.97
foau1240	flh	Abawiri	139.2	-3.05
abaz1241	abq	Abaza	42.0	44.25
abkh1244	abk	Abkhaz	41.2	43.06
abua1245	aah	Abu' Arapesh	142.9	-3.46
abui1241	abz	Abui	124.6	-8.31
achi1257	ace	Acehnese	96.6	3.91
achu1248	acu	Achuar-Shiwiar	-77.3	-2.83
acol1236	ach	Acoli	32.5	3.58

Below shows the first ten segments and the first ten featural values of a modified version of PanPhon (Mortensen et al. 2016).

ipa	syl	son	cons	cont	delrel	lat	nas	strid	voi	sg
q	-1	-1	1	-1	-1	-1	-1	0	-1	-1
q̣	-1	-1	1	-1	-1	-1	-1	0	-1	-1
q̠	-1	-1	1	-1	-1	-1	-1	0	-1	-1

ipa	syl	son	cons	cont	delrel	lat	nas	strid	voi	sg
ɢ	-1	-1	1	-1	-1	-1	-1	0	1	-1
ɢ̬	-1	-1	1	-1	-1	-1	-1	0	1	-1
ɢ̲	-1	-1	1	-1	-1	-1	-1	0	1	-1
qː	-1	-1	1	-1	-1	-1	-1	0	-1	-1
ɢː	-1	-1	1	-1	-1	-1	-1	0	1	-1
ʈ	-1	-1	1	-1	-1	-1	-1	0	-1	-1
ʈ̥	-1	-1	1	-1	-1	-1	-1	0	-1	-1

Measuring the distance between sequences

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

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

As an example, the table below shows the featural values of /t/ and /p/. The *gap* column is the gap between each of /t/'s featural value and each of /p/'s corresponding featural value. The sum of these gaps is 5, which is the Saporta distance between /t/ and /p/.

Feature	t	p	gap
syl	-1	-1	0
son	-1	-1	0
cons	1	1	0
cont	-1	-1	0
delrel	-1	-1	0
lat	-1	-1	0
nas	-1	-1	0
strid	0	0	0
voi	-1	-1	0
sg	-1	-1	0
cg	-1	-1	0
ant	1	1	0
cor	1	-1	2

Feature	t	p	gap
distr	-1	0	1
lab	-1	1	2
hi	-1	-1	0
lo	-1	-1	0
back	-1	-1	0
round	-1	-1	0
velaric	-1	-1	0
tense	0	0	0
long	-1	-1	0

Although the Saporta distance is the distance between two segments and not sequences, I will apply it to measure the distance between sequences.

In order to measure the distance between two sequences of different length. I assign different "positions" to each sequence. As the maximal length of all sequences is six, a sequence of only one segment has six positions within these six slots. For example, the sequence /p/ would have the following six positions:

Slot1	Slot2	Slot3	Slot4	Slot5	Slot6
p					
	p				
		p			
			p		
				p	
					p

In the first row, where /p/ is assigned the leftmost position, the features of Slot 1 (lab1, voi1 …) are equivalent to the phonological features of /p/. In the remaining five slots (lab2, voi2, … lab6, voi6 …), the corresponding featural values are 0.

For the sequence /pl/, five positions are assigned within six slots:

Slot1	Slot2	Slot3	Slot4	Slot5	Slot6
p	l				
	p	l			
		p	l		
			p	l	

Slot1	Slot2	Slot3	Slot4	Slot5	Slot6
				p	l

For the sequence /spl/, four positions are assigned within six slots:

Slot1	Slot2	Slot3	Slot4	Slot5	Slot6
s	p	l			
	s	p	l		
		s	p	l	
			s	p	l

In order to compare the distance between /pl/ and /spl/, /pl/ in each of the five positions is mapped onto /spl/ in each of the four positions (5 * 4 = 20 comparisons). The table below shows the first five comparisons.

Comparison	Slot1	Slot2	Slot3	Slot4	Slot5	Slot6
1	s	p	l			
1	p	l				
2	s	p	l			
2		p	l			
3	s	p	l			
3			p	l		
4	s	p	l			
4				p	l	
5	s	p	l			
5					p	l

Among the five comparisons, the second comparison yields the minimal distance between /spl/ and /pl/, as /p/ is compared to /p/, /l/ is compared to /l/, and /s/ is compared to zero. Thus, the second comparison is chosen as the distance between /spl/ and /pl/.[1] The Distance between /pl/ and /spl/ is thus the sum of the Saporta distance between /s/ and zero values, the Saporta distance between /p/ and /p/, and the Saporta Saporta distance between /l/ and /l/. As the distance between /p/ and /p/ and the distance between /l/ and /l/ are zero, the distance between /spl/ and /pl/ is effectively the distance between /s/ and zero values.

As an example, below are shown the sequences that are the most similar to /pl/.

[1] I thank Huisu Yun for suggesting this idea to me.

Sequence.x	Sequence.y	Distance
pl	pl	0
pl	bl	2
pl	pʰl	2
pl	p̬l	2
pl	ⁿpl	2
pl	pˀl	2
pl	plʲ	2
pl	fl	3
pl	p͡fl	3
pl	plʏ	4
pl	m̥l	4
pl	b̬l	4
pl	bʰl	4
pl	pʰl̥	4
pl	b̤l	4
pl	pz	4
pl	ⁿbl	4
pl	bˀl	4
pl	blʲ	4
pl	ɓl	4

As another example, below are shown the sequences that are the most similar to /ia/.

Sequence.x	Sequence.y	Distance
ia	ia	0
ia	iæ	0
ia	ie	2
ia	ea	2
ia	ɪa	2
ia	iã	2
ia	ia:	2
ia	i:a	2
ia	ĩa	2
ia	ɨa	2
ia	eæ	2
ia	iæ:	2
ia	i:æ	2

Sequence.x	Sequence.y	Distance
ia	iɑ	3
ia	iɛ	4
ia	iːe	4
ia	aa	4
ia	ee	4
ia	ii	4
ia	ya	4

Measuring the distance between lects

In this section, I will show how I measure the phonological distance between two lects.

I calculate the distance between two lects within the same category (onset, nucleus, or coda). The distance between the onset/nucleus/coda sequences of two lects is defined as follows. Let M1 and M2 be the matrices representing the phonological feature values of the onset/nucleus/coda sequences of lect 1 and lect 2, respectively. Let MD be the distance matrix between M1 and M2. The distance between the onset/nucleus/coda forms of lect 1 and 2 is the average value of the minimum values of rows or columns of MD, whichever is higher.

As an example, suppose that lect A allows three onset sequences, /p m t/, and lect B two onset sequences, /p t/. The comparison between A and B is shown in the table below. The last column and the last row shows the minimum value of each row and each column, respectively. The average value of the minimum column (the comparison from A to B) is $\frac{0+6+0}{3} = 2$, whereas the average value of the minimum row (the comparison from B to A) is 0. The bigger value of these two is selected. Thus, the onset distance between A and B is 2.

A_B	p	t	min
p	0	5	0
m	6	11	6
t	5	0	0
min	0	0	NA

Using this formula, I calculate the distance of onset, nucleus, and coda of each pair of lects. Below shows the first ten pairs.

Lect_vs_Lect	O	N	C
A'ou vs. A'ou	0.00	0.00	0.00
A'ou vs. Akajeru	3.83	4.72	5.59
A'ou vs. Amdo Tibetan	8.61	20.80	5.00
A'ou vs. Angami Naga	2.42	10.36	13.00
A'ou vs. Ao Naga	3.83	10.20	5.00
A'ou vs. Archi	4.39	20.24	22.98
A'ou vs. Arvanitika Albanian	12.62	9.72	17.98
A'ou vs. Asho Chin	10.03	20.24	12.18
A'ou vs. Assamese	9.81	10.36	10.87
A'ou vs. Asturian-Leonese-Cantabrian	4.35	2.16	6.41

Measuring the distance of tones

Next, I calculate the distance between the tonality of each pair of lects.

The distance between tonality is defined as the Canberra distance between the numbers of tonemes of two lects. Let T1 be the number of tonemes lect 1 has, and T2 the number of tonemes lect 2 has. The distance betwen lect 1 and lect 2 is $\frac{|T1-T2|}{(T1+T2)}$.

For example, Burmese has 3 tonemes, whereas Yue Chinese has 6. The tonal distance (T) between two lects is thus $\frac{|3-6|}{(3+6)} = \frac{1}{3}$.

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

First, I count the number of tones in each lect. Below shows the first ten lects.

Lect	T
A'ou	4
Akajeru	0
Amdo Tibetan	0
Angami Naga	5
Ao Naga	3
Archi	0
Arvanitika Albanian	0
Asho Chin	3
Assamese	0
Asturian-Leonese-Cantabrian	0

I calculate the Canberra distance between the numbers of tones of each pair of lects. Below shows the first ten pairs.

Lect_vs_Lect	T
A'ou vs. A'ou	0.000
A'ou vs. Akajeru	2.000
A'ou vs. Amdo Tibetan	2.000
A'ou vs. Angami Naga	0.222
A'ou vs. Ao Naga	0.286
A'ou vs. Archi	2.000
A'ou vs. Arvanitika Albanian	2.000
A'ou vs. Asho Chin	0.286
A'ou vs. Assamese	2.000
A'ou vs. Asturian-Leonese-Cantabrian	2.000

I join segmental distance with tonal distance and normalize the four distances. Below shows the first ten rows.

Lect_vs_Lect	O	N	C	T
A'ou vs. A'ou	-1.563	-1.532	-1.187	-1.088
A'ou vs. Akajeru	-0.876	-0.451	-0.536	1.020
A'ou vs. Amdo Tibetan	-0.019	3.229	-0.604	1.020
A'ou vs. Angami Naga	-1.129	0.840	0.328	-0.854
A'ou vs. Ao Naga	-0.876	0.803	-0.604	-0.787
A'ou vs. Archi	-0.776	3.101	1.491	1.020
A'ou vs. Arvanitika Albanian	0.699	0.693	0.908	1.020
A'ou vs. Asho Chin	0.235	3.101	0.232	-0.787
A'ou vs. Assamese	0.196	0.840	0.080	1.020
A'ou vs. Asturian-Leonese-Cantabrian	-0.783	-1.037	-0.440	1.020

Measuring the overall distance

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

Lect_vs_Lect	Distance
A'ou vs. A'ou	0.00
A'ou vs. Akajeru	2.55
A'ou vs. Amdo Tibetan	5.46
A'ou vs. Angami Naga	2.86
A'ou vs. Ao Naga	2.52
A'ou vs. Archi	5.80
A'ou vs. Arvanitika Albanian	4.35
A'ou vs. Asho Chin	5.18
A'ou vs. Assamese	3.84
A'ou vs. Asturian-Leonese-Cantabrian	2.42

Below shows the five closest lects per each sample lect.

Lect	1	2	3	4	5
A'ou	Bugan	Min Bei Chinese	Gan Chinese	Hakka Chinese	Jinyu Chinese
Akajeru	Asturian-Leonese-Cantabrian	Santali	Sinhala	Bih	Forest Enets
Amdo Tibetan	Sri Lanka Malay	Dongxiang	Mangghuer	Tangam	Western Magar
Angami Naga	Narua	Eastern Kayah	Zauzou	Daohua	Western Xiangxi Miao
Ao Naga	Thado Chin	Cosao	Pela	Zeme Naga	Bolyu
Archi	Bagvalal	Avar	Uighur	Chuvash	Burushaski
Arvanitika Albanian	Pite Saami	Dutch	Cypriot Arabic	Assamese	Macedonian
Asho Chin	Nocte Naga	Darma	Kado	Zaiwa	Japanese
Assamese	Dutch	Kathmandu Valley Newari	Chitwania Tharu	Wambule	Gata'
Asturian-Leonese-Cantabrian	Bih	Akajeru	Mongghul	Tshangla	Maithili
Atong (India)	Standard Malay	Kelantan-Pattani Malay	Tundra Nenets	Jarawa (India)	Bonan
Avar	Archi	Uighur	Bagvalal	Chuvash	Kirghiz

Lect	1	2	3	4	5
Baba Malay	Nepali	Nihali	Korra Koraga	Sindhi	Hindi
Badaga	Tundra Nenets	Kelantan-Pattani Malay	Jarawa (India)	South Azerbaijani	Ravula
Bagvalal	Archi	Avar	Burushaski	Chuvash	Uighur
Bantawa	Tangam	Western Magar	Dongxiang	Mangghuer	Koi
Basque	Daman-Diu Portuguese	Assamese	Gata’	Welsh	Bunan
Betta Kurumba	Godwari	Koi	Kathmandu Valley Newari	Southern Amami-Oshima	Jennu Kurumba
Bih	Asturian-Leonese-Cantabrian	South Wa	Khmu	Tshangla	Mon
Bisu	Chut	Kado	Hills Karbi	Thai	Koireng
Biyo	Bugan	Gan Chinese	Wu Chinese	Min Bei Chinese	Hui Chinese
Bodo-Mech	Moyon	Thai	Chut	Bisu	Darma
Bolyu	Cosao	Cao Miao	Sadu	Thado Chin	Central Hongshuihe Zhuang
Bonan	Tundra Nenets	Atong (India)	Standard Malay	Jarawa (India)	Kelantan-Pattani Malay
Bugan	Min Bei Chinese	Cosao	Gan Chinese	Biyo	A’ou
Bulo Stieng	Moken	Nganasan	Dhivehi	Sora	Jarawa (India)
Bunan	Kashmiri	Daman-Diu Portuguese	Assamese	Kathmandu Valley Newari	Pite Saami
Burmese	Kado	Yerong-Southern Buyang	Iu Mien	Western Parbate Kham	Jiongnai Bunu
Burushaski	Sadri	Uighur	Gilaki	Kirghiz	Chuvash

Lect	1	2	3	4	5
Cao Miao	Central Hongshuihe Zhuang	Bolyu	Yongbei Zhuang	Northern Pinghua	Cosao
Central Chong	Chut	Thai	Bisu	Southern Jinghpaw	Koireng
Central Hongshuihe Zhuang	Cao Miao	Northern Pinghua	Bolyu	Cosao	Min Nan Chinese
Central Khmer	Mlabri	Laven	Assamese	Kadar	Mongghul
Chak	Zbu	Japhug	Nocte Naga	Asho Chin	Wambule
Chintang	Atong (India)	Udihe	Kelantan-Pattani Malay	Semelai	Bonan
Chitwania Tharu	Wambule	Italian	Yakkha	Assamese	Kathmandu Valley Newari
Chut	Thai	Bisu	Central Chong	Hills Karbi	Southern Jinghpaw
Chuvash	Uighur	Kirghiz	Evenki	Sadri	Kazakh
Cosao	Bolyu	Pela	Thado Chin	Bugan	Sadu
Cypriot Arabic	Arvanitika Albanian	Russian	Dutch	Neo-Mandaic	Welsh
Daai Chin	Eastern Panjabi	Yerong-Southern Buyang	Western Parbate Kham	Koireng	Central Chong
Dagur	Kelantan-Pattani Malay	Tundra Nenets	Atong (India)	Vach-Vasjugan	Jarawa (India)
Daman-Diu Portuguese	Basque	Bunan	Gata’	Estonian Swedish	Portuguese
Dandami Maria	Gilaki	Jennu Kurumba	Pite Saami	Korra Koraga	Hindi
Danish	Hungarian	Macedonian	Gurani	Pite Saami	Purik-Sham-Nubra

Lect	1	2	3	4	5
Daohua	Enu	Zauzou	Narua	Angami Naga	Wuding-Luquan Yi
Darma	Nocte Naga	Kado	Japanese	Asho Chin	Moyon
Dhivehi	Sora	Marathi	Bulo Stieng	Russia Buriat	Moken
Dongxiang	Mangghuer	Western Magar	Tangam	Bantawa	Amdo Tibetan
Dutch	Assamese	Pite Saami	Arvanitika Albanian	Sindhi	Kathmandu Valley Newari
E	Mulam	Yerong-Southern Buyang	Iu Mien	Jiongnai Bunu	Yongbei Zhuang
Eastern Katu	Mon	Khmu	Kashmiri	South Wa	Kadar
Eastern Kayah	Angami Naga	Narua	Daohua	Zauzou	Khams Tibetan
Eastern Magar	Southern Pashto	Chitwania Tharu	Wambule	Kathmandu Valley Newari	Italian
Eastern Panjabi	Daai Chin	Koireng	Rabha	Western Parbate Kham	Chut
Enu	Daohua	Zauzou	Narua	Wuding-Luquan Yi	Biyo
Ersu	Shixing	Wuding-Luquan Yi	Southern Pumi	Enu	Khams Tibetan
Estonian Swedish	Maithili	Welsh	Dutch	Daman-Diu Portuguese	Gata’
Evenki	Kirghiz	Tatar	Northern Yukaghir	Tuvinian	Sadri
Forest Enets	Sinhala	Akajeru	Santali	Peripheral Mongolian	West Yugur
French	Arvanitika Albanian	Bunan	Maithili	Estonian Swedish	Russian
Galo	Tibetan	Solu-Khumbu Sherpa	Moyon	Kyerung	Lao

Lect	1	2	3	4	5
Gan Chinese	Hui Chinese	Min Bei Chinese	Wu Chinese	Hakka Chinese	Jinyu Chinese
Gata’	Assamese	Dutch	Chitwania Tharu	Italian	Kathmandu Valley Newari
German	Danish	Russian	Ostfränkisch	Cypriot Arabic	Hungarian
Gilaki	Burushaski	Jennu Kurumba	Pite Saami	Sindhi	Dandami Maria
Godoberi	Southern Pumi	Rabha	Modern Greek	Mulam	Miyako
Godwari	Betta Kurumba	Koi	Southern Amami-Oshima	Malayalam	Italian
Gurani	Hungarian	Sindhi	Nihali	Baba Malay	Gilaki
Hakka Chinese	Gan Chinese	Min Nan Chinese	Min Bei Chinese	Tsat	Hui Chinese
Halh Mongolian	Pite Saami	Bagvalal	Peripheral Mongolian	Avar	Uighur
Hills Karbi	Koireng	Rabha	Bisu	Southern Jinghpaw	Thai
Hindi	Jennu Kurumba	Sindhi	Nihali	Korra Koraga	Sholaga
Hinuq	Marathi	Dhivehi	Russia Buriat	Sora	Bulo Stieng
Hokkaido Ainu	Standard Malay	Atong (India)	Kelantan-Pattani Malay	Jarawa (India)	Tundra Nenets
Hui Chinese	Gan Chinese	Wu Chinese	Min Bei Chinese	Sadu	Hakka Chinese
Hungarian	Pite Saami	Danish	Gilaki	Gurani	Macedonian
Icelandic	Mlabri	Wambule	Betta Kurumba	Modern Greek	Eastern Katu
Italian	Chitwania Tharu	Wambule	Kathmandu Valley Newari	Koi	Khmu

Lect	1	2	3	4	5
Iu Mien	Mulam	Yerong-Southern Buyang	E	Jiongnai Bunu	Yongbei Zhuang
Japanese	Darma	Nocte Naga	Kado	Moyon	Kyerung
Japhug	Purik-Sham-Nubra	Laz	Amdo Tibetan	Macedonian	Pnar
Jarawa (India)	Tundra Nenets	Atong (India)	Standard Malay	Ravula	Kelantan-Pattani Malay
Jejueo	Korean	Malacca-Batavia Portuguese Creole	Santali	Nganasan	Bulo Stieng
Jennu Kurumba	Hindi	Sindhi	Muduga	Korra Koraga	Sholaga
Jinyu Chinese	Gan Chinese	Hui Chinese	Min Bei Chinese	Bugan	Tsat
Jiongnai Bunu	Mulam	Iu Mien	E	Yerong-Southern Buyang	Vietnamese
Kadar	Tshangla	Eastern Katu	Khmu	Mon	Modern Greek
Kado	Nocte Naga	Bisu	Darma	Western Parbate Kham	Yerong-Southern Buyang
Kashmiri	Eastern Katu	Jennu Kurumba	Mon	Kodava	Sholaga
Kathmandu Valley Newari	Italian	Koi	Wambule	Assamese	Chitwania Tharu
Katso	Enu	Daohua	Biyo	Wuding-Luquan Yi	Zauzou
Kazakh	Kirghiz	Tatar	Uighur	Evenki	Sadri
Kelantan-Pattani Malay	Tundra Nenets	Standard Malay	Atong (India)	Vach-Vasjugan	South Azerbaijani
Khams Tibetan	Sadu	Narua	Daohua	Zeme Naga	Biyo
Khmu	Mon	Eastern Katu	Bih	Koi	South Wa
Kirghiz	Tatar	Kazakh	Evenki	Uighur	Sadri

Lect	1	2	3	4	5
Kodava	Muduga	Malavedan	Sholaga	Jennu Kurumba	Korra Koraga
Koi	Southern Amami-Oshima	Kathmandu Valley Newari	Italian	Wambule	Mon
Koireng	Hills Karbi	Zeme Naga	Rabha	Bisu	Southern Jinghpaw
Korean	Jejueo	Malacca-Batavia Portuguese Creole	Nganasan	Bulo Stieng	Santali
Korra Koraga	Sindhi	Nihali	Jennu Kurumba	Baba Malay	Nepali
Kyerung	Solu-Khumbu Sherpa	Yue Chinese	Thai	Northeastern Thai	Lao
Lao	Northeastern Thai	Sadu	Yue Chinese	Southern Jinghpaw	Wu Chinese
Laven	Pnar	Malayalam	Sedang	Modern Greek	Yakkha
Laz	Leh Ladakhi	Purik-Sham-Nubra	Japhug	Macedonian	Arvanitika Albanian
Leh Ladakhi	Purik-Sham-Nubra	Sindhi	Jennu Kurumba	Pite Saami	Sri Lanka Malay
Macedonian	Pite Saami	Purik-Sham-Nubra	Hungarian	Arvanitika Albanian	Dutch
Maithili	Estonian Swedish	Gata’	Assamese	Basque	Welsh
Malacca-Batavia Portuguese Creole	Korean	Jejueo	Santali	Russia Buriat	Dhivehi
Malavedan	Muduga	Kodava	Sholaga	Sri Lanka Malay	Jennu Kurumba
Malayalam	Wambule	Chitwania Tharu	Assamese	Southern Pashto	Yakkha

Lect	1	2	3	4	5
Mandarin Chinese	Cosao	Thado Chin	Zeme Naga	Pela	Sadu
Mangghuer	Dongxiang	Western Magar	Tangam	Bantawa	Amdo Tibetan
Maonan	Zaiwa	Kado	Nocte Naga	Vietnamese	Kyerung
Marathi	Dhivehi	Russia Buriat	Sora	Bulo Stieng	Hinuq
Min Bei Chinese	Gan Chinese	Bugan	Hakka Chinese	Hui Chinese	Jinyu Chinese
Min Nan Chinese	Hakka Chinese	Northern Pinghua	Tsat	Hui Chinese	Gan Chinese
Miyako	Western Parbate Kham	Oki-No-Erabu	E	Rabha	Southern Pumi
Mlabri	Italian	Central Khmer	Chitwania Tharu	Kathmandu Valley Newari	Khmu
Modern Greek	Italian	Assamese	Chitwania Tharu	Kadar	Khmu
Moken	Bulo Stieng	Dhivehi	Nganasan	Sora	Jarawa (India)
Mon	Khmu	Eastern Katu	South Wa	Kashmiri	Koi
Mongghul	Kadar	Khmu	Yakkha	Koi	Modern Greek
Moyon	Bodo-Mech	Galo	Solu-Khumbu Sherpa	Darma	Kyerung
Muduga	Malavedan	Sholaga	Kodava	Jennu Kurumba	Vaagri Booli
Mulam	Iu Mien	E	Yerong-Southern Buyang	Jiongnai Bunu	Yongbei Zhuang
Narua	Angami Naga	Daohua	Zauzou	Enu	Sani
Naukan Yupik	Tundra Nenets	Kelantan-Pattani Malay	Atong (India)	Standard Malay	Hokkaido Ainu

Lect	1	2	3	4	5
Neo-Mandaic	Welsh	Cypriot Arabic	Dutch	Estonian Swedish	Daman-Diu Portuguese
Nepali	Baba Malay	Nihali	Korra Koraga	Sindhi	Hindi
Nganasan	Bulo Stieng	Sora	Moken	Dhivehi	Russia Buriat
Nihali	Nepali	Sindhi	Baba Malay	Korra Koraga	Western Magar
Nocte Naga	Darma	Kado	Asho Chin	Japanese	Maonan
Northeastern Thai	Lao	Sadu	Yue Chinese	Kyerung	Wu Chinese
Northern Pinghua	Central Hongshuihe Zhuang	Cao Miao	Min Nan Chinese	Min Bei Chinese	Hui Chinese
Northern Yukaghir	Southern Yukaghir	Ravula	Tuvinian	Evenki	Jarawa (India)
Oki-No-Erabu	Western Parbate Kham	Kado	Koireng	Sangkong	Darma
Ostfränkisch	German	Russian	Hungarian	Estonian Swedish	Danish
Pa-Hng	Solu-Khumbu Sherpa	Yue Chinese	Kyerung	Sangkong	Northeastern Thai
Pela	Cosao	Bolyu	Thado Chin	Sadu	Bugan
Peripheral Mongolian	Sakha	West Yugur	Santali	Kashmiri	Tamil
Pite Saami	Jennu Kurumba	Gilaki	Sindhi	Dutch	Hungarian
Pnar	Sri Lanka Malay	Laven	Pite Saami	Malayalam	Khmu
Portuguese	Tshangla	Modern Greek	Malayalam	South Wa	Daman-Diu Portuguese
Pu-Xian Chinese	Hui Chinese	Gan Chinese	Min Bei Chinese	Min Nan Chinese	Biyo

Lect	1	2	3	4	5
Purik-Sham-Nubra	Leh Ladakhi	Macedonian	Nihali	Sindhi	Dutch
Rabha	Koireng	Hills Karbi	Bisu	Chut	Central Chong
Ravula	Northern Yukaghir	Jarawa (India)	Southern Yukaghir	Tuvinian	Tundra Nenets
Russia Buriat	Dhivehi	Sora	Marathi	Bulo Stieng	Sakha
Russian	Cypriot Arabic	Arvanitika Albanian	Welsh	Dutch	German
Sadri	Tatar	Uighur	Kirghiz	Evenki	Southern Yukaghir
Sadu	Bolyu	Thado Chin	Hui Chinese	Cosao	Pela
Sakha	West Yugur	Bulo Stieng	Russia Buriat	Dhivehi	Sora
Sangkong	Yerong-Southern Buyang	Southern Jinghpaw	Zaiwa	Bisu	Kado
Sani	Western Xiangxi Miao	Narua	Angami Naga	Western Muya	Khams Tibetan
Santali	Russia Buriat	Dhivehi	West Yugur	Sinhala	Sora
Sedang	Khmu	Laven	Eastern Katu	Assamese	Gata’
Semelai	Standard Malay	Kelantan-Pattani Malay	Jarawa (India)	Atong (India)	Tundra Nenets
Shixing	Southern Pumi	Biyo	Bugan	A’ou	Min Bei Chinese
Sholaga	Muduga	Jennu Kurumba	Kodava	Malavedan	Vaagri Booli
Sichuan Yi	Western Muya	Angami Naga	Japanese	Khams Tibetan	Eastern Kayah
Sindhi	Korra Koraga	Nihali	Hindi	Jennu Kurumba	Muduga
Sinhala	Santali	Forest Enets	Akajeru	Russia Buriat	West Yugur

Lect	1	2	3	4	5
Solu-Khumbu Sherpa	Kyerung	Galo	Yue Chinese	Tibetan	Lao
Sora	Dhivehi	Russia Buriat	Bulo Stieng	Nganasan	Marathi
South Azerbaijani	Kelantan-Pattani Malay	Tundra Nenets	Standard Malay	Atong (India)	Jarawa (India)
South Wa	Mon	Bih	Tshangla	Eastern Katu	Khmu
Southeast Pashayi	Zeme Naga	Koireng	Thado Chin	Hills Karbi	Southern Jinghpaw
Southern Amami-Oshima	Koi	Kathmandu Valley Newari	Italian	Mon	Khmu
Southern Jinghpaw	Hills Karbi	Thai	Koireng	Lao	Bisu
Southern Pashto	Malayalam	Modern Greek	Italian	Chitwania Tharu	Kathmandu Valley Newari
Southern Pumi	Shixing	A'ou	Biyo	Yerong-Southern Buyang	Bugan
Southern Yukaghir	Northern Yukaghir	Ravula	Tuvinian	Sadri	Evenki
Sri Lanka Malay	Malavedan	Sindhi	Muduga	Pnar	Pite Saami
Standard Malay	Atong (India)	Kelantan-Pattani Malay	Tundra Nenets	Jarawa (India)	Hokkaido Ainu
Tamil	Dhivehi	Marathi	Tuvinian	Bulo Stieng	Sakha
Tangam	Bantawa	Vaagri Booli	Dongxiang	Sholaga	Muduga
Tatar	Kirghiz	Kazakh	Uighur	Sadri	Evenki
Thado Chin	Cosao	Bolyu	Sadu	Pela	Zeme Naga
Thai	Chut	Southern Jinghpaw	Bisu	Kyerung	Hills Karbi
Tibetan	Galo	Solu-Khumbu Sherpa	Kyerung	Yue Chinese	Lao
Toda	Dagur	Betta Kurumba	Vach-Vasjugan	Peripheral Mongolian	Eastern Katu

Lect	1	2	3	4	5
Tsat	Gan Chinese	Hakka Chinese	Min Nan Chinese	Hui Chinese	Bugan
Tshangla	Sora	Dhivehi	Russia Buriat	South Wa	Marathi
Tundra Nenets	Kelantan-Pattani Malay	Standard Malay	Jarawa (India)	Vach-Vasjugan	Atong (India)
Tuvinian	Ravula	Tundra Nenets	Jarawa (India)	Kelantan-Pattani Malay	South Azerbaijani
Udihe	Tundra Nenets	Kelantan-Pattani Malay	Jarawa (India)	Chintang	Vach-Vasjugan
Uighur	Tatar	Kirghiz	Chuvash	Sadri	Kazakh
Vaagri Booli	Muduga	Tangam	Jennu Kurumba	Sholaga	Sindhi
Vach-Vasjugan	Tundra Nenets	Kelantan-Pattani Malay	Standard Malay	South Azerbaijani	Jarawa (India)
Vietnamese	Thai	Iu Mien	Jiongnai Bunu	Maonan	Kado
Wambule	Chitwania Tharu	Italian	Kathmandu Valley Newari	Malayalam	Koi
Welsh	Estonian Swedish	Dutch	Neo-Mandaic	Basque	Cypriot Arabic
West Yugur	Sakha	Santali	Bulo Stieng	Russia Buriat	Peripheral Mongolian
Western Magar	Nihali	Mangghuer	Dongxiang	Bantawa	Tangam
Western Muya	Angami Naga	Sani	Narua	Western Xiangxi Miao	Khams Tibetan
Western Parbate Kham	Kado	Oki-No-Erabu	Yerong-Southern Buyang	Mulam	Chut
Western Xiangxi Miao	Sani	Angami Naga	Narua	Western Muya	Eastern Kayah
Wu Chinese	Hui Chinese	Gan Chinese	Biyo	Sadu	Bolyu

Lect	1	2	3	4	5
Wuding-Luquan Yi	Daohua	Zauzou	Enu	Katso	Shixing
Wutunhua	Dongxiang	Tangam	Mangghuer	Bantawa	Western Magar
Yakkha	Chitwania Tharu	Wambule	Malayalam	Italian	Pnar
Yerong-Southern Buyang	Yongbei Zhuang	Iu Mien	Mulam	E	Sangkong
Yongbei Zhuang	Yerong-Southern Buyang	Cao Miao	Bolyu	Central Hongshuihe Zhuang	Mulam
Yue Chinese	Kyerung	Northeastern Thai	Solu-Khumbu Sherpa	Lao	Tibetan
Zaiwa	Maonan	Sangkong	Southern Jinghpaw	Darma	Kado
Zauzou	Daohua	Angami Naga	Enu	Narua	Eastern Kayah
Zbu	Japhug	Asho Chin	Nocte Naga	Bodo-Mech	Darma
Zeme Naga	Koireng	Thado Chin	Mandarin Chinese	Southeast Pashayi	Hills Karbi

Clustering the lects

Based on these distances, I perform two analyses: Multidimensional scaling and k-means clustering, in order to cluster similar lects together and detect areal patterns.

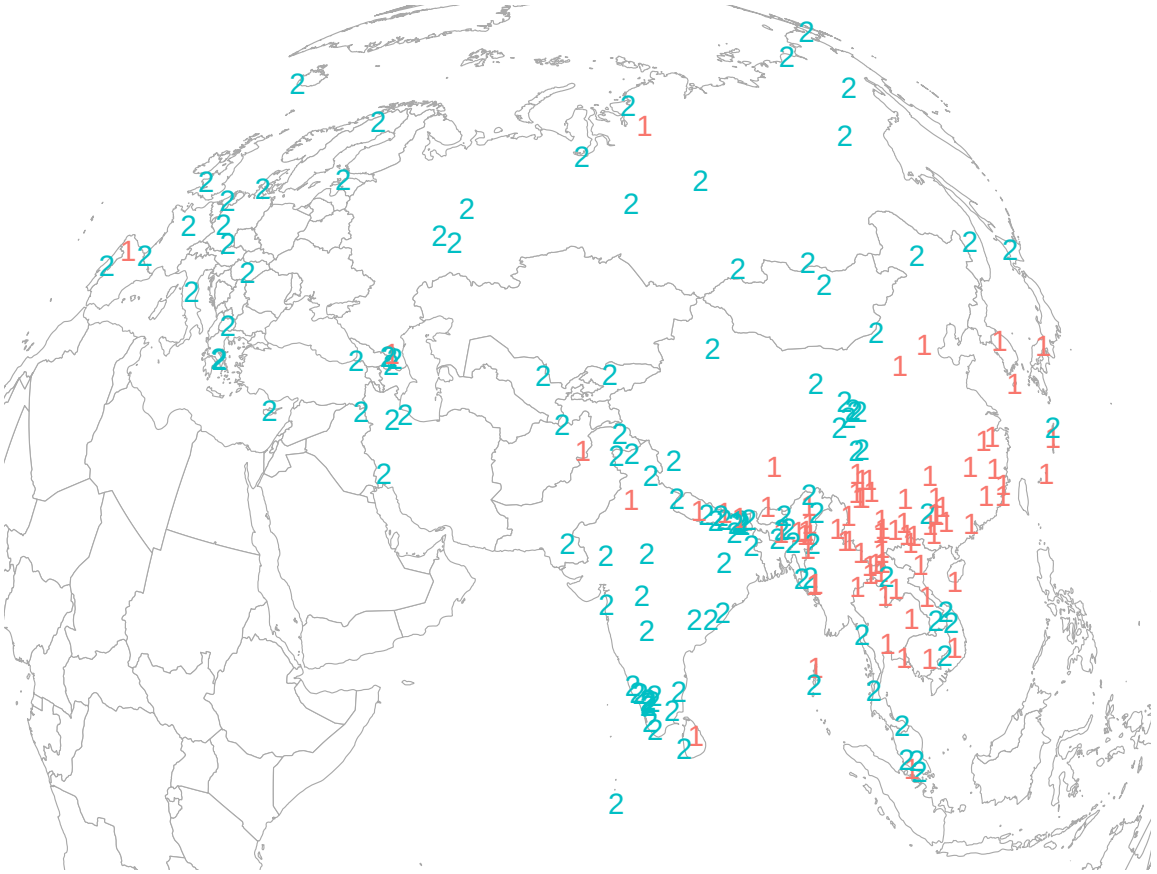
I conduct multidimensional scaling based on the phonological distances, the number of dimensions being maximal, i. e. the number of sample lects minus one. Below shows the first ten pairs of lects and the first five dimensions.

Lect	V1	V2	V3	V4	V5
A'ou	-3.131	2.011	-0.477	-0.312	-0.324
Akajeru	-1.419	2.021	-1.493	0.962	0.030

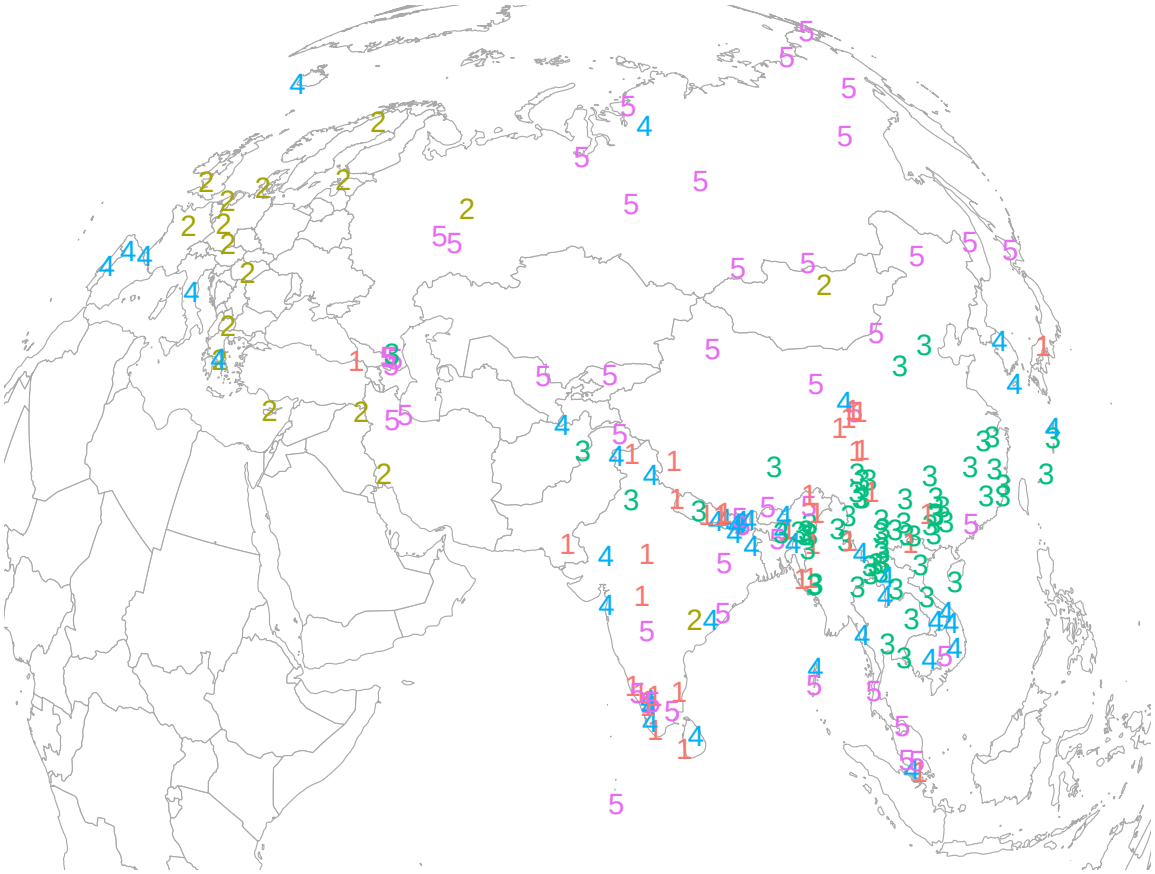
Lect	V1	V2	V3	V4	V5
Amdo Tibetan	1.296	-0.853	0.792	0.631	-0.200
Angami Naga	-2.090	-0.909	1.150	0.684	-0.624
Ao Naga	-1.973	0.246	-0.491	-0.757	0.343
Archi	1.530	-1.318	-1.511	-0.456	-0.043
Arvanitika Albanian	1.732	1.576	0.224	-0.030	-0.584
Asho Chin	0.721	-0.543	1.707	-1.277	0.431
Assamese	1.087	1.171	0.117	0.337	-0.245
Asturian-Leonese-Cantabrian	-1.729	3.132	-1.313	1.049	-0.145

Next, based on the multidimensional scaling, I will perform k-means clustering of two, five, and ten clusters, in order to see if the clusters formed based on phonological distance consist of areally close lects.

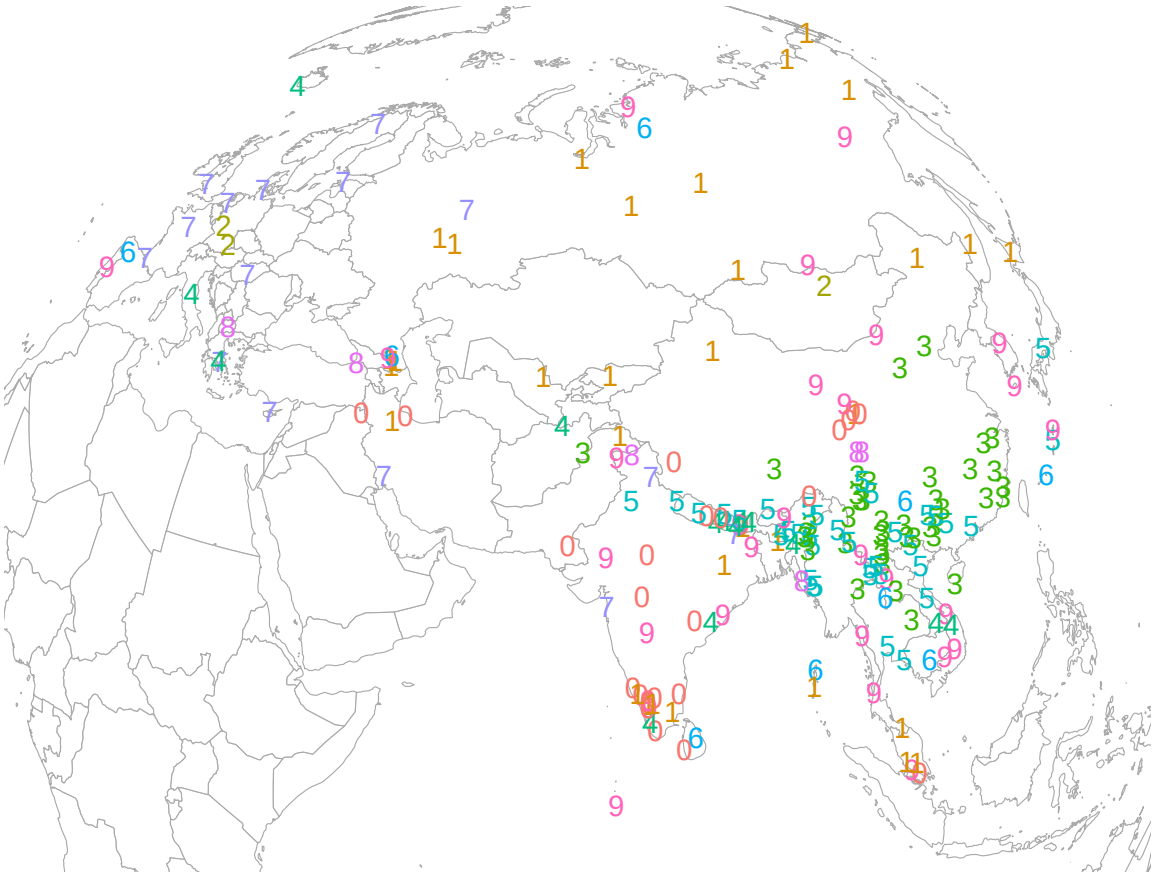
I assign the two clusters on the map of Eurasia, each integer in different colors representing different clusters. We see that East and Southeast Asia are distinct from the rest of the macroarea.



Below shows the five clusters on the map.



Below shows the ten clusters on the map.



From the visualized k-means clustering, we can make the following observations: Phonological clusters also tend to form geographical clusters. Lects in Northeast Asia, Qinghai-Gansu, Mainland South East Asia, South Asia, and Europe tend to form clusters. Thus, we can confirm that these form phonological areas.

Testing the correlation between phonological and geographical distances

I will also test the following hypothesis: Geographical distance correlates with phonological distance. That is, geographically closer lects also tend to be phonologically similar.

I calculate the geographical distances between two columns of coordinates. I leave out pairs of lects that belong to the same family, as lects belonging to the same family tend to be phonologically similar (by inheritance) and also geographically closer. Below shows the first ten geographical distances.

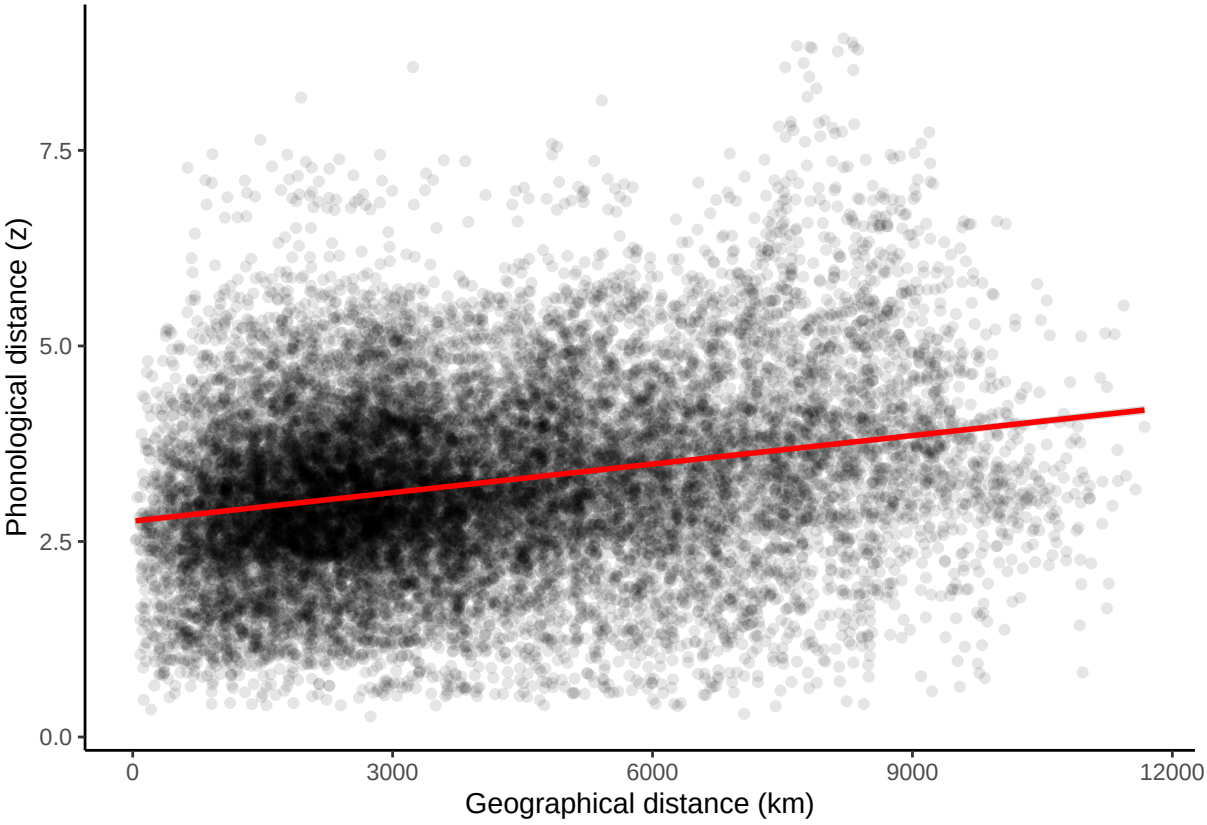
Lect_vs_Lect	Kilometers
A'ou vs. Akajeru	2029
A'ou vs. Amdo Tibetan	1001
A'ou vs. Angami Naga	1202
A'ou vs. Ao Naga	1143
A'ou vs. Archi	5562
A'ou vs. Arvanitika Albanian	7594
A'ou vs. Asho Chin	1390
A'ou vs. Assamese	1452
A'ou vs. Asturian-Leonese-Cantabrian	9644
A'ou vs. Atong (India)	1527

I conduct linear regression between geographical distance and phonological distances.

```
##
## Call:
## lm(formula = Distance ~ Kilometers, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.365 -0.696 -0.070  0.646  5.412
##
```

```
## Coefficients:
##              Estimate Std. Error t value           Pr(>|t|)
## (Intercept) 2.76027663 0.01589725     174 <0.0000000000000002 ***
## Kilometers  0.00012159 0.00000348      35 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.12 on 18349 degrees of freedom
## Multiple R-squared:  0.0625, Adjusted R-squared:  0.0625
## F-statistic: 1.22e+03 on 1 and 18349 DF,  p-value: <0.0000000000000002
```

Finally, I visualize the linear regression.



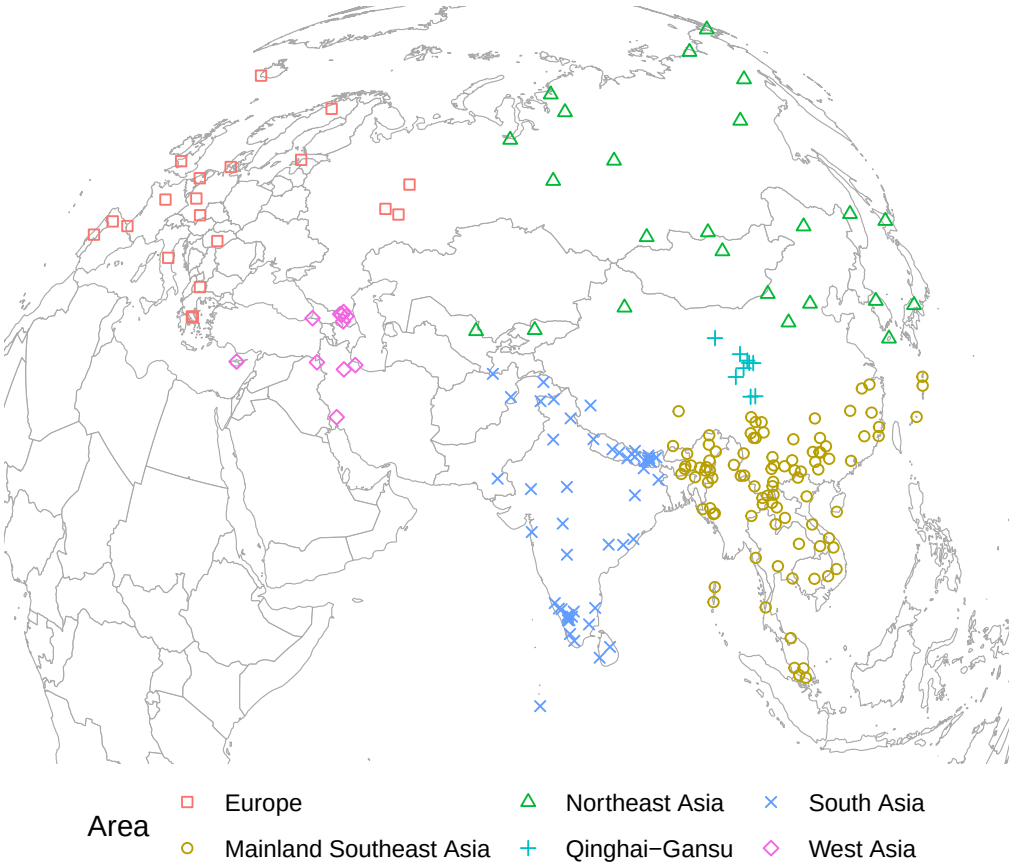
We see that there is a correlation between the geographical distance and the phonological distance between lects that are not genealogically related. We can thus make the following conclusion: Geographically close lects tend to phonologically converge.

Naive Bayes Classifier

As a follow-up study, I will examine how well machine learning predicts the the area of a lect given its phonological distance from other lects. For example, based on how similar German is to other Eurasian lects, can we predict that it is spoken in Europe?

I first divide the Eurasian lects into six different regions solely based on their geographical coordinates: Northeast Asia, Mainland Southeast Asia, Qinghai-Gansu, South Asia, West Asia, and Europe.

The map visualizes the lects in the predefined six areas.



The goal is train a model based on phonological distance to see how well it predicts which one of these six areas a lect is spoken.

I train the Naive Bayes Classifier based on half of the lects and their distance from other lects. First, I divide the sample in half by each area. (The proportion of the areas is thus equal in the halved sample.) Then I train the classifier in the first half and test it on the other half. Below shows the first ten predictions.

Lect	Prediction	Area	Correct
A'ou	Mainland Southeast Asia	Mainland Southeast Asia	TRUE
Akajeru	Mainland Southeast Asia	Mainland Southeast Asia	TRUE
Amdo Tibetan	South Asia	Qinghai-Gansu	FALSE
Angami Naga	Mainland Southeast Asia	Mainland Southeast Asia	TRUE
Archi	West Asia	West Asia	TRUE
Arvanitika Albanian	South Asia	Europe	FALSE
Asho Chin	South Asia	Mainland Southeast Asia	FALSE

Lect	Prediction	Area	Correct
Assamese	South Asia	Mainland Southeast Asia	FALSE
Atong (India)	Northeast Asia	Mainland Southeast Asia	FALSE
Baba Malay	South Asia	Mainland Southeast Asia	FALSE

I perform the same training and testing, but training with the latter half as the training and testing the former half. Below shows the first ten predictions.

Lect	Prediction	Area	Correct
Ao Naga	Mainland Southeast Asia	Mainland Southeast Asia	TRUE
Asturian-Leonese-Cantabrian	Mainland Southeast Asia	Europe	FALSE
Avar	West Asia	West Asia	TRUE
Badaga	Northeast Asia	South Asia	FALSE
Bagvalal	West Asia	West Asia	TRUE
Basque	Europe	Europe	TRUE
Betta Kurumba	South Asia	South Asia	TRUE
Bih	South Asia	Mainland Southeast Asia	FALSE
Bisu	Mainland Southeast Asia	Mainland Southeast Asia	TRUE
Bolyu	Mainland Southeast Asia	Mainland Southeast Asia	TRUE

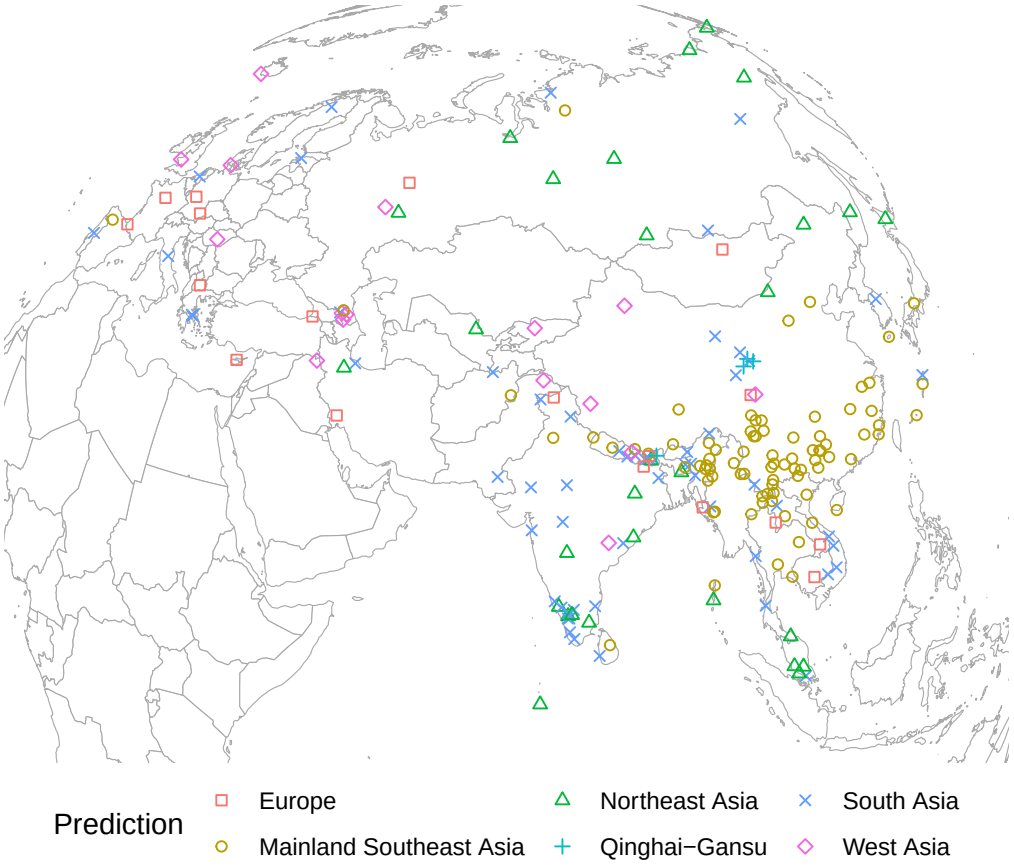
Below is the confusion matrix and the related statistics, based on the combination of the two halves of prediction. The accuracy of the predictions is significantly higher than the No Information Rate ($p < 0.001$) The Kappa value also shows that the model successfully predicts the areas to a moderate degree.

```
##      Accuracy         Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##      0.576190      0.413868       0.506294       0.643904       0.452381
## AccuracyPValue  McnemarPValue
##      0.000211           NaN
```

The F1 values of individual classes show that the model predicts some areas better than others, namely Northeast Asia, Mainland Southeast Asia, and South Asia. This may be partly because these areas have more sample lects than others.

Class	F1
Europe	0.316
Mainland Southeast Asia	0.775
Northeast Asia	0.453
Qinghai-Gansu	0.462
South Asia	0.486
West Asia	0.296

Below is the visualization of the lects by their predicted areas. Note that the "wrongly" predicted areas may not be necessarily "wrong" in the sense that the initial division of Eurasia may not fully reflect real phonological areas and the wrong predictions may actually be "correct" in the sense that they in fact belong to the "wrongly" predicted areas.



5.3 Summary

This chapter has reached the second goal of this thesis: Measuring the phonological distance between Eurasian lects using Phonotacticon 1.0. Importantly, the distance measuring is based on the entirety of the phonotactic data available in Phonotacticon, except for the tonal qualities (as only the number of tones was used to calculate the distance between tonal inventories). Clustering the lects together based on phonological distance shows that phonologically close lects also tend to be genealogically close and that some of the clusters correspond with previously suggested linguistic areas, as we have discussed in Section 2.3. Moreover, machine learning based on the phonological distances can predict the area of each lect to a moderate degree.

Chapter 6

Conclusion

In this thesis, I have shown how I have built a phonotactic database (§3), used that database to generate visualizations showing the diverse areal patterns of Eurasian phonology (§4), and also used it to measure the phonological distances between the sample lects (§5). The results show that phonological distances correlate with geographical distance in the Eurasian macroarea, geographically closer lects being phonologically more similar. The areal clusters based on phonological distance confirms the linguistic areas introduced in Section 2.3.2: Northeast Asia, Mainland Southeast Asia, South Asia, and Europe. Thus, we have evidence that these regions form phonological areas, if not linguistic areas including convergence in other domains as well.

Needless to say, the purpose of Phonotacticon does not end here. It can be used for a wide range of purposes, such as producing different visualizations based on a wide range of phonotactic features, or calculate the phonological distance between lects with different methodologies. Moreover, as the database is still in progress, I plan to complete Phonotacticon 2.0 in the following years, including lects from all the macroareas. I will also engage in measuring phonological distance and detecting phonological areas using other methods. I hope that Phonotacticon will, beyond this dissertation, function as a fruitful database for diverse interests of many phonologists and typologists.

References

Abbi, Anvita (2018). "Echo formations and expressives in South Asian languages". In: *Non-prototypical reduplication*. Ed. by Aina Urdze. De Gruyter, pp. 1–34. DOI: 10.1515/9783110599329-001.

Afendras, Evangelos A. (1970). "Quantitative distinctive feature typologies and a demonstration of areal convergence". In: *ITL-International Journal of Applied Linguistics* 9.1, pp. 49–81. DOI: 10.1075/itl.9.05afe.

Anderson, Cormac, Tiago Tresoldi, Simon J. Greenhill, Robert Forkel, Russell D Gray, and Johann-Mattis List (2021). "Measuring variation in phoneme inventories". Version 1. In: DOI: 10.21203/rs.3.rs-891645/v1.

Anderson, Gregory D. S. (2006). "Towards a typology of the Siberian linguistic area". In: *Linguistic areas: Convergence in historical and typological perspective*. Ed. by Yaron Matras, April McMahon, and Nigel Vincent. Palgrave Macmillan London, pp. 266–300. DOI: 10.1057/9780230287617_11.

Arsenault, Paul Edmond (2012). "Retroflex consonant harmony in South Asia". PhD thesis. University of Toronto.

Avram, Andrei (1964). "Sur la typologie phonologique quantitative [On the quantiative phonological typology]". In: *Revue roumaine de Linguistique* IX, pp. 131–134.

Benjamin, Geoffrey (1976). "An outline of Temiar grammar". In: *Austroasiatic Studies Part 1*. Ed. by Philip N. Jenner, Laurence C. Thompson, and Stanley Starosta. Vol. 13. Oceanic Linguistics Special Publications. University of Hawai'i Press, pp. 129–187.

Bisang, Walter (2006). "Linguistic areas, language contact and typology: Some implications from the case of Ethiopia as a linguistic area". In: *Linguistic areas: Convergence in historical and typological perspective*. Ed. by Yaron Matras, April McMahon, and Nigel Vincent. Palgrave Macmillan London, pp. 75–98. DOI: 10.1057/9780230287617_4.

Blevins, Juliette (2002). "Notes on sources of Yurok glottalized consonants". In: *Proceedings of the Meeting of the Hokan-Penutian Workshop: Survey of California and Other Indian Languages*. Ed. by Laura Buszard-Welcher and Leanne Hinton. Vol. 11. University of California, pp. 1–18.

— (2017). "Areal sound patterns: From perceptual magnets to stone soup". In: *The Cambridge handbook of areal linguistics* 5587.

Brown, Cecil H. (2013). “Finger and Hand”. In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/chapter/130.

Cardoso, Hugo C. (2009). “The Indo-Portuguese language of Diu”. PhD thesis. Universiteit van Amsterdam.

Chen 陳, Naixiong 乃雄 (1988). “Wutunhua yinxi 五屯話音系 [The sound system of Wutun speech]”. In: *Minzu yuwen* 民族語文 3, pp. 1–10.

Chirikba, Viacheslav A. (2008). “The problem of the Caucasian Sprachbund”. In: *From linguistic areas to areal linguistics*. Ed. by Pieter Muysken. Vol. 90. Studies in language companion series. John Benjamins Publishing Company, pp. 25–93.

Chirkova, Katia, James N. Stanford, and Dehe Wang (2018). “A long way from New York City: Socially stratified contact-induced phonological convergence in Ganluo Ersu (Sichuan, China)”. In: *Language Variation and Change* 30.1, pp. 109–145. DOI: 10.1017/S095439451700028X.

Clements, George (1990). “The role of the sonority cycle in core syllabification”. In: *Papers in laboratory phonology*. Ed. by John Kingston and Mary Beckman. Vol. 1, pp. 283–333. DOI: 10.1017/CBO9780511627736.017.

Comrie, Bernard (2007). “Areal typology of Mainland Southeast Asia: What we learn from the WALS maps”. In: *MANUSYA: Journal of Humanities* 10.3, pp. 18–47.

de Sousa, Hilário (2015). “The Far Southern Sinitic languages as part of Mainland Southeast Asia”. In: *Languages of Mainland Southeast Asia: The State of the Art*. Ed. by Nick James Enfield and Bernard Comrie. De Gruyter Mouton, pp. 356–440. DOI: 10.1515/9781501501685-009.

Dingemanse, Mark (2019). ““Ideophone” as a comparative concept”. In: *Ideophones, mimetics, and expressives*. Ed. by Kimi Akita and Prashant Pardeshi. John Benjamins Publishing Company, pp. 13–33.

Dixon, R. M. W. and Alexandra Y. Aikhenvald (2003). “Word: A typological framework”. In: *Word: A Cross-linguistic Typology*. Ed. by R. M. W. Dixon and Alexandra Y. Aikhenvald. Cambridge University Press, pp. 1–41. DOI: 10.1017/CBO9780511486241.002.

Do, Youngah and Ryan Ka Yau Lai (2021). “Accounting for lexical tones when modeling phonological distance”. In: *Language* 97.1, e39–e67.

Donohue, Mark (2013). “Who inherits what, when?: Toward a theory of contact, substrates, and superimposition zones”. In: *Language Typology and Historical Contingency: In honor of Johanna Nichols*. Ed. by Balthasar Bickel, Lenore A. Grenoble, David A. Peterson, and Alan Timberlake. John Benjamins, pp. 219–240.

Doornenbal, Marius (2009). “A grammar of Bantawa: Grammar, paradigm tables, glossary and texts of a Rai language of Eastern Nepal”. PhD thesis. Rijksuniversiteit te Leiden.

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

Dwyer, Arienne (2008). “Tonogenesis in southeastern Monguor”. In: *Lessons from documented endangered languages*. Typological studies in language 78. Ed. by K. David Harrison, David S. Rood, and Arienne Dwyer, pp. 111–128.

— (2013). “Tibetan as a dominant Sprachbund language: Its interactions with neighboring languages”. In: *The third international conference on the Tibetan language*. Trace Foundation, pp. 258–280.

Ebihara 海老原, Shiho 志保 (2019). *Amudo chibettogo bunpō* アムド・チベット語文法 *[Amdo Tibetan grammar]*. Hitsuji shobō ひつじ書房.

Eden, S. Elizabeth (2018). “Measuring phonological distance between languages”. PhD thesis. University College London.

Emeneau, Murray B. (1956). “India as a lingustic area”. In: *Language* 32.1, pp. 3–16.

— (1969). “Onomatopoetics in the Indian linguistic area”. In: *Language* 45.2, pp. 274–299.

Enfield, Nick James (2018). *Mainland Southeast Asian Languages: A Concise Typological Introduction*. Cambridge University Press.

Evans, Nicholas (2019). “Linguistic divergence under contact”. In: *Historical Linguistics 2015: Selected papers from the 22nd International Conference on Historical Linguistics, Naples, 27-31 July 2015*. Ed. by Michela Cennamo and Claudia Fabrizio. Vol. 348. Current Issues in Linguistic Theory. John Benjamins Publishing Company, pp. 564–591. DOI: 10.1075/cilt.348.26eva.

Flikeid, Karin and Wladyslaw Cichocki (1987). “Application of dialectometry to Nova Scotia Acadian French dialects: Phonological distance.” In: *Papers from the Annual Meetings of the Atlantic Provinces Linguistic Association (PAMAPLA)* 11, pp. 59–74.

François, Alexandre (2011). “Social ecology and language history in the northern Vanuatu linkage: A tale of divergence and convergence”. In: *Journal of Historical Linguistics* 1.2, pp. 175–246.

Georg, Stefan (2008). “Yeniseic languages and the Siberian linguistic area”. In: *Evidence and counter-evidence: Essays in Honour of Frederik Kortlandt*. Ed. by Alexander Lubotsky, Jos Schaeken, and Jeroen Wiedenhof. Vol. 2: General linguistics. Studies in Slavic and General Linguistics. Brill, pp. 151–168. DOI: 10.1163/9789401206365_011.

Gerner, Matthias (2013). *A Grammar of Nuosu*. Vol. 64. Mouton Grammar Library. De Gruyter Mouton, p. 543.

Gowda, K. S. Gurubasave (1968). “Descriptive analysis of Soliga”. PhD thesis. Deccan College.

Grossman, Eitan, Elad Eisen, Dmitry Nikolaev, and Steven Moran (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, pp. 5316–5322.

Gruzdeva, Ekaterina (1998). *Nivkh*. Vol. 111. Languages of the World/Materials. Lincom.

Gut, Ulrike (2009). *Introduction to English phonetics and phonology*. Vol. 1. Textbooks in English Language and Linguistics (TELL). Peter Lang GmbH.

Hammarström, Harald and Mark Donohue (2014). “Some principles on the use of macro-areas in typological comparison”. In: *Language Dynamics and Change* 4.1, pp. 167–187.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank (2021). *Glottolog 4.4*. Max Planck Institute for Evolutionary Anthropology. DOI: 10.5281/zenodo.4761960.

Harnud, Huhe and Xuewen Zhou (2021). “On the relation between the similarity of the acoustic distribution patterns of vowels and the language closeness”. In: *International Journal of Anthropology and Ethnology* 5.1, pp. 1–13.

Haspelmath, Martin (1998). “How young is Standard Average European?” In: *Language Sciences* 20.3, pp. 271–287.

— (2001). “The European linguistic area: Standard Average European”. In: *Language Typology and Language Universals / Sprachtypologie und sprachliche Universalien / La typologie des langues et les universaux linguistiques*. Ed. by Martin Haspelmath. Vol. 2. De Gruyter Mouton. Chap. 107, pp. 1492–1510. DOI: 10.1515/9783110194265-044.

Hölzl, Andreas (2018). *A Typology of Questions in Northeast Asia and beyond: An Ecological Perspective*. Studies in Diversity Linguistics. Language Science Press.

Huang 黃, Yan 燕 (2007). “Gu nilaimu zi zai xiandai hanyu fangyan zhong de fenhun qingkuang 古泥來母字在現代漢語方言中的分混情況 [The conditions of mixed and separate Ni Lai initial consonant in contemporary dialect]”. In: *Journal of Suzhou University* 宿州學院學報 22.5, pp. 64–67.

Itô, Junko and Armin Mester (1999). “The phonological lexicon”. In: *The handbook of Japanese linguistics*. Ed. by Natsuko Tsujimura. Blackwell Publishers Ltd., pp. 62–100. DOI: 10.1002/9781405166225.ch3.

Iwasaki, Shoichi (2013). *Japanese*. Ed. by Theodora Bynon, David C. Bennett, and Masayoshi Shibtani. Revised. Vol. 17. London Oriental and African Language Library. John Benjamins Publishing Company.

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

Jenny, Mathias and San San Hnin Tun (2016). *Burmese: A comprehensive grammar*. Routledge.

Joseph, Brian D. (2020). “Language contact in the Balkans”. In: *The handbook of language contact*. Ed. by Raymond Hickey. 2nd ed. John Wiley & Sons, Ltd, pp. 537–549. DOI: 10.1002/9781119485094.ch27.

Kahn, Daniel (1976). “Syllable-based generalizations in English phonology”. PhD thesis. Massachusetts Institute of Technology.

Kang, Yoonjung and Sungwoo Han (2013). “Tonogenesis in early contemporary Seoul Korean: A longitudinal case study”. In: *Lingua* 134, pp. 62–74.

Kučera, Henry and George K. Monroe (1968). *A comparative quantitative phonology of Russian, Czech, and German*. Mathematical Linguistics and Automatic Language Processing 4. American Elsevier Publishing Company.

Kühl, Karoline and Kurt Braunmüller (2014). "Linguistic stability and divergence". In: *Stability and divergence in language contact: Factors and mechanisms*. Ed. by Kurt Braunmüller, Steffen Höder, and Karoline Kühl. Vol. 16. Studies in Language Variation. John Benjamins Publishing Company, pp. 13–38.

Li, Xia, Jinfang Li, and Yongxian Luo (2014). *A grammar of Zoulei (Southwest China)*. Peter Lang.

Macklin-Cordes, Jayden L, Claire Bowern, and Erich R Round (2021). "Phylogenetic signal in phonotactics". In: *Diachronica* 38.2, pp. 210–258.

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

— (2013a). "Syllable structure". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/chapter/12.

— (2013b). "Tone". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/chapter/13.

Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino (2013). "LAPSyd: Lyon-Albuquerque phonological systems database". In: *Interspeech 2013*. International Speech Communication Association (ISCA). DOI: 10.21437/interspeech.2013-660.

Malaia, Evie A. and Ronnie B. Wilbur (2020). "Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 11.1, e1518.

Malotki, Ekkehart (1983). *Hopi time: A linguistic analysis of the temporal concepts in the Hopi language*. Vol. 20. Trends in Linguistics. Studies and Monographs [TiLSM]. De Gruyter Mouton. DOI: 10.1515/9783110822816.

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

Meakins, Felicity and Rob Pensalfini (2021). "Holding the mirror up to converted languages: Two grammars, one lexicon". In: *International Journal of Bilingualism* 25.2, pp. 425–457.

Mielke, Jeff (2008). *The emergence of distinctive features*. Oxford Series in Typology and Linguistic Theory. Oxford University Press.

Moral, Dipankar (1997). "North-east India as a linguistic area". In: *Mon-Khmer Studies* 27, pp. 43–54.

Moran, Steven, Eitan Grossman, and Annemarie Verkerk (2021). "Investigating diachronic trends in phonological inventories using BDPROTO". In: *Language Resources and Evaluation* 55.1, pp. 79–103.

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

Mortensen, David R., Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and 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*, pp. 3475–3484.

Nikolaev, Dmitry (2018). "The Database of Eurasian Phonological Inventories: A research tool for distributional phonological typology". In: *Linguistics Vanguard* 4.1.

— (2019). "Areal dependency of consonant inventories". In: *Language Dynamics and Change* 9.1, pp. 104–126.

Nugteren, Hans and Marti Roos (1996). "Common vocabulary of the Western and Eastern Yugur languages: The Turkic and Mongolic loanwords". In: *Acta Orientalia Academiae Scientiarum Hungaricae* 49.1/2, pp. 25–91.

Oganian, Yulia and Edward F. Chang (2019). "A speech envelope landmark for syllable encoding in human superior temporal gyrus". In: *Science Advances* 5.11, eaay6279. DOI: 10.1126/sciadv.aay6279.

Ohala, John J. (1990). "There is no interface between phonology and phonetics: A personal view". In: *Journal of phonetics* 18.2, pp. 153–171.

Okada, Hideo (1991). "Japanese". In: *Journal of the International Phonetic Association* 21.2, pp. 94–96.

Panov, Vladimir (2020). "Final particles in Asia: Establishing an areal feature". In: *Linguistic Typology* 24.1, pp. 13–70.

Patrie, James (1982). *The genetic relationship of the Ainu Language*. University of Hawaii Press.

Peyraube, Alain (2017). "The case system in three Sinitic languages of the Qinghai-Gansu linguistic area". In: *Languages and genes in northwestern China and adjacent regions*. Ed. by Dan Xu and Hui Li. Springer, pp. 121–139. DOI: 10.1007/978-981-10-4169-3_8.

Pike, Kenneth L. (1947). "On the phonemic status of English diphthongs". In: *Language* 23.2, pp. 151–159.

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

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. URL: https://www.R-project.org.

Refsing, Kirsten (1986). *The Ainu language: The morphology and syntax of the Shizunai dialect*. Aarhus University Press.

Riad, Tomas (2013). *The phonology of Swedish*. The Phonology of the World's Languages. Oxford University Press.

Sandman, Erika (2016). "A grammar of Wutun". PhD thesis. University of Helsinki.

Saporta, Sol (1955). "Frequency of consonant clusters". In: *Language* 31.1, p. 25. DOI: 10.2307/410889.

Schapper, Antoinette Schapper, Lila San Roque, and Rachel Hendery (2016). "Tree, firewood and fire in the languages of Sahul". In: *The Lexical Typology of Semantic Shifts*. Ed. by Päivi Juvonen and Maria Koptjevskaja-Tamm. De Gruyter Mouton, pp. 355–422. DOI: 10.1515/9783110377675-012.

Schiering, René, Balthasar Bickel, and Kristine A. Hildebrandt (2010). "The prosodic word is not universal, but emergent". In: *Journal of Linguistics* 46.3, pp. 657–709.

Sidwell, Paul and Mathias Jenny (2021a). "Introduction". In: *The languages and linguistics of Mainland Southeast Asia: A comprehensive guide*. Ed. by Paul Sidwell and Mathias Jenny. De Gruyter Mouton, pp. 1–20. DOI: 10.1515/9783110558142-001.

— (2021b). *The Languages and Linguistics of Mainland Southeast Asia: A comprehensive guide*. De Gruyter Mouton. DOI: 10.1515/9783110558142.

Simeon, George (1969). "Hokkaido Ainu phonemics". In: *Journal of the American Oriental Society*, pp. 751–757.

Slater, Keith W. (2003). *A grammar of mangghuer: A Mongolic language of China's Qinghai-Gansu sprachbund*. Routledge Curzon Asian Linguistics Series. Routledge Curzon, p. 382.

Szeto, Pui Yiu and Chingduang Yurayong (2021). "Sinitic as a typological sandwich: Revisiting the notions of Altaicization and Taicization". In: *Linguistic Typology* 25.3, pp. 551–599.

Tamura, Suzuko (2000). *The Ainu language*. 1st ed. Vol. 2. ICHEL Linguistic Studies. Sanseido.

Thomason, Sarah Grey (2000). "Linguistic areas and language history". In: *Languages in Contact*. Ed. by Dicky Gilbers, John Nerbonne, and Jos Schaeken. Vol. 28. Studies in Slavic and General Linguistics, pp. 311–327.

Trubetzkoy, Nikolai (1928). "Proposition 16". In: *Actes du Premier Congrès International de Linguistes : à La Haye, du 10-15 avril 1928*. Ed. by Antoine Meillet. Uilgeversmaatschappij, pp. 17–18.

Vajda, Edward J. (2008). "The languages of Siberia". In: *Language and Linguistics Compass* 3.1, pp. 424–440. DOI: 10.1111/j.1749-818x.2008.00110.x.

van der Hulst, Harry (2017). "Phonological typology". In: *The Cambridge Handbook of Linguistic Typology*. Ed. by Alexandra Y. Aikhenvald and R. M. W. Dixon. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, pp. 39–77. DOI: 10.1017/9781316135716.002.

Vittrant, Alice and Justin Watkins, eds. (2019). *The Mainland Southeast Asia linguistic area*. De Gruyter Mouton. DOI: 10.1515/9783110401981.

Whitman ホイットマン, John ジョン (2016). “Tōhoku ajia gengo chiikino ichi dzukeni mukete 東北アジア言語地域の位置付けに向けて [On the Northeast Asia as a linguistic area]”. In: 国語研プロジェクトレビュー= *NINJAL Project Review* 6, pp. 69–82.

Whorf, Benjamin Lee (1944). “The relation of habitual thought and behavior to language”. In: *ETC: A Review of General Semantics* 1.4, pp. 197–215.

Wiese, Richard (2000). *The phonology of German*. The Phonology of the World’s Languages. Oxford University Press.

Wu, Hugjiltu (2003). “Bonan”. In: *The Mongolic languages*. Ed. by Juha Janhunen. Routledge Language Family Series. Routledge, pp. 325–345.

Xie, Yihui, Christophe Dervieux, and Emily Riederer (2020). *R Markdown cookbook*. ISBN 9780367563837. Chapman and Hall/CRC. URL: https://bookdown.org/yihui/rmarkdown-cookbook.

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

Yi, Han Gyol, Matthew K. Leonard, and Edward F. Chang (2019). “The encoding of speech sounds in the superior temporal gyrus”. In: *Neuron* 102.6, pp. 1096–1110.

Yu, Mengxia, Ce Mo, You Li, and Lei Mo (2015). “Distinct representations of syllables and phonemes in Chinese production: Evidence from fMRI adaptation”. In: *Neuropsychologia* 77, pp. 253–259.

Yurayong, Chingduang and Pui Yiu Szeto (2020). “Altaicization and de-Altaicization of Japonic and Koreanic”. In: *International Journal of Eurasian Linguistics* 2.1, pp. 108–148.

Zakaria, Muhammad (2018). “A grammar of Hyow”. PhD thesis. Nanyang Technological University.

Zhou, Chenlei (2020). “Case markers and language contact in the Gansu-Qinghai linguistic area”. In: *Asian Languages and Linguistics* 1.1, pp. 168–203.