PHONOLOGICAL AREAS IN EURASIA

IAN JOO

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies

Phonological areas in Eurasia

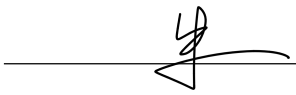
Ian Joo

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

February 2024

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 (linguistic varieties) have converged into similar phonological patterns. In order to compute the distribution of phonological areas in Eurasia, I have built Phonotacticon 1.0, a cross-linguistic database that contains 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, namely Europe, South Asia, Qinghai-Gansu, Northeast Asia, and 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 am also greatly indebted to Dr. Yao Yao, the chairwoman of the board of examiners, and Prof. Ian Maddieson and Prof. Mattis List, the two external examiners, for their thorough reviews of this dissertation.

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 516 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.
- The appendix shows the R script used to obtain the statistical results and visualizations employed throughout 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

The main goal of this thesis is to detect phonological areas in Eurasia. A phonological area is defined as a geographical space inhabited by speakers of lects that have phonologically converged into one another. What, then, is 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 from 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.

Klein et al. (2020) observe that phonological patterns are generally transferred from one lect to another only within the transferred vocabulary, although suprasegmental patterns (such as tones) may be an exception to this rule. For example, Sino-Japanese initial /r-/ did not cause native Japonic words to develop initial /r-/; rather, it remained within the Sinitic loanwords and other strata that permit /r-/. This also applies to, according to Klein et al. (2020), phonotactic constraints: in Hindi-Urdu, complex onsets and codas are only found in loans from Sanskrit, English, Persian, and Arabic. These loans did not cause native Hindi-Urdu words to develop complex onset or coda.

Loanwords are not the only origin of external import of phonological patterns, however. Boretzky (1991) illustrates two cases which he claims to be phonological transfer without lexical transfer: the diphthongization of Czech /iː/ and /uː/ due to the influence of Old High German and the vowel centralization in Kalderash Romani motivated by Romanian. (The latter is cited by Klein et al. (2020) as an exception to their generalization.) Blevins (2002) also reports the emergence of ejectives in Yurok, an Algic lect spoken in Northwestern California. Yurok has the ejective consonants /t͡ʃʼ kʼ kʷʼ pʼ tʼ/, which are not found in its sister lect Wiyok. Blevins (2017, p. 96) argue that although /tʼ t͡ʃʼ/ 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 convergence) 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, I 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 translation "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 protolect 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 will be shown in Section 5.2.5.9). They have never been proposed as belonging to the same language family, however, which must be proven not by typological similarities but the classical methods of historical comparative linguistics based on the cognacy of basic vocabulary. Also, given the huge geographical distance between the two peoples' inhabited areas and the absence of any known historical contact between the two peoples, it is highly unlikely 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).

I should also add that this criterion for areahood is the **similarity due to** contact and not **contact per se**. In other words, the occurrence of language contact is not enough and only after the contact leads to convergence can it contribute to areahood. This is because, as discussed in Section 2.2, a contact does not always lead to convergence: it can lead to divergence or no contact-induced changes at all. Thus, a region not qualifying as a linguistic area with shared linguistic similarities does not necessarily imply that no contact has happened there. Likewise, the historical evidence of contact between multiple linguistic groups does not automatically prove that their lects form a linguistic area.

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 2013c). Thus, tone is a non-unique areal feature of the Mainland Southeast Asia.

What make a feature areal, then, is not necessarily its uniqueness, 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 difficult 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 (§2.3.2.1). 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 Caucasus

Caucasus is the mountainous isthmus between the Caspian Sea and the Black Sea, shared by the present-day states of Russia, Georgia, Turkey, Armenia, and Azerbaijan. It is one of the most linguistically diverse regions in the world, especially its northern part (Comrie 2008). Caucasus is home to three language families unique to it: Kartvelian (South Caucasian), Abkhaz-Adyge (Northwest Caucasian), and Nakh-Daghestanian (Northeast Caucasian). It is also populated by the speakers of some Turkic and Indo-European lects, such as Azerbaijani and Armenian. Whether the genealogically diverse lects of the Caucasus form a high enough

degree of homogeneity to make Caucasus a linguistic area remains unclear and understudied.

Catford's (1977) survey of fifteen linguistic features among four Caucasian language groups - Nakh, Dagestanian, Kartvelian, and Northwest Caucasian - finds that only three features are shared by all four groups: uvular consonants, glottalic consonants, and ergative constructions. Tuite (1999) yet criticizes that these three features are not adequate for a linguistic area to be established. He only accepts glottalic consonants as a valid areal feature, as it is shared not only by the families unique to Caucasus but by the neighboring Indo-European and Turkic lects as well. Uvular consonants, on the other hand, are far too common across Eurasia to be an areal characteristic. Moreover, he argues that the types of ergativity in different Caucasian families are so fundamentally distinct in nature that ergativity cannot be viewed as a genuine commonality.

Comrie (2008) suggests that endogamy might have played a role in maintaining the divergent patterns in Caucasus. When speakers tend to marry members of their own ethnic group, linguistic diffusion across different lect groups may be slowed down, leading to divergence rather than convergence. Balanovsky et al. (2011), based on Y-chromosomal data, observe that four linguistic groups inhabiting different areas of the Caucasus - Nakh, Dagestanian, Ossetic, and Abkhazo-Adyghian - also form four different genetic clusters (Turkic speakers were not included in their study, due to their recent origin). Comrie (2008) also suggests that the mountainous terrain might also be relevant for the linguistic diversity, although he is uncertain about the causative relationship.

On the other hand, Chirikba (2008) and Daniel and Lander (2011) provide some linguistic features typical of the Caucasian area, such as:

- **Rich consonant inventory**. Most Caucasian lects have a huge consonant inventory. Among the thirteen Caucasian lects sampled in the World Atlas of Linguistic Structures, ten are classified as having a large consonant inventory (34 or more), two (Georgian and Eastern Armenian) as having a moderately large one (26 to 33), and only one (Azerbaijani) as having one of average size (19 to 25) (Maddieson 2013a).

- **Agglutinative morphology**. Caucasian morphology is mostly agglutinative, having little morphemic fusion and instead encoding grammatical information by affixation. This areal feature has led Azerbaijani (Turkic), Ossetic (Indo-European), and Armenian (Indo-European) to lose their declensional characters and converge into agglutinative morphology, according to Chirikba (2008, p. 51) and Daniel and Lander (2011, p. 131).

- **Left-branching SOV word order**. While Caucasian lects have flexible word order, the basic word order remains subject-object-verb. It is unclear, however, whether this feature is distinctive enough to be areally specific, as most of Eurasia excluding Europe and Southeast Asia is dominated by SOV lects (Dryer 2013).

Vogt (1988) argues that the rich nominal case system, shared by Georgian, Armenian, Ossetic, and North(west) Caucasian, is a result of the linguistic convergence in Caucasus. While he does not mention Turkic, Azerbaijani also has six cases, although the rich case system is a general feature of the Turkic family. Daniel and Lander (2011) also view nominal case as "typical of the Caucasus" (p. 133), Nakh-Dagestanian being the exception for having up to two core cases only, nominative and oblique. Both Vogt (1988) and Daniel and Lander (2011) highlight Ossetic, whose numerous cases are neither found often in other Iranic lects nor mostly inherited from Indo-European.

Whether these features are numerous and "heavy" enough to validate the areality of the Caucasus is unclear. Regardless of whether Caucasus is a linguistic area or not, all previous studies agree that there is some amount of commonalities between the lects of the Caucasus and the question remains whether that amount is sufficient, which can only be answered in quantitative terms. This thesis (§5–5.3) may be able to fill in that gap from a data-driven perspective.

2.3.2.2 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).

2.3.2.3 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 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 (2018, §3.5) also mentions Siberia and Qinghai-Gansu (cf. §2.3.2.5) as commonly proposed subareas of Northeast Asia. But he remains skeptical of the areahood of these two regions, arguing that Siberian and Qinghai-Gansu features are too typologically common to qualify the two regions as linguistic areas. 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 more 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.

2.3.2.5 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, I 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. /hd͡zɑ/ '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 characteristics of Mangghuer include the retroflex consonants /ʂ t͡ʂ t͡ʂʰ ɻ/, which are typical to northern Sinitic. The simplicity of its 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 (H. Wu 2003).

As mentioned in Section 2.3.2.4, 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 show 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.

2.3.2.6 South Asia

South Asia, largely equivalent to the Indian subcontinent, is a linguistic area dominated by Indo-Aryan (branch of Indo-European) lects in the north and Dravidian lects in the south, while also home to many Sino-Tibetan and Mundaic (branch of 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 throughout 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’

Note that reduplication is a very common morphological strategy used not only in South Asia but in most areas of the world, Europe being rather exceptional for not using it extensively (Rubino 2013). What makes South Asian echo formation special is then not its reduplicative morphology but its shared set of semantic functions, which may not be served by reduplications in other areas. Li and Ponsford’s (2018) survey of 108 lects shows that while certain meanings are commonly expressed by reduplication, such as iterativity and intensivity, certain meanings are expressed by reduplication in a relatively small number of lects, such as

randomness or negation. It is thus important to state that reduplication per se is not an areal feature of South Asia but reduplication in the semantic range as illustrated by Abbi (2018).

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* ਲਸ ਲਸ '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).

2.3.3 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 Phonological databases

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

2.4.1 UCLA Phonological Segment Inventory Database (UPSID)

The UCLA Phonological Segment Inventory Database, or UPSID (Maddieson 2009), which is accessible at `web.phonetik.uni-frankfurt.de/upsid.html`), and was released in 1984, is

the oldest phonological database that is currently available online. It consists of the phonemic inventory of 451 lects across the world. Although it is not without limitations, such as only containing segmental information and not tones, UPSID remains a useful phonological database at present.

2.4.2 The Database of Eurasian Phonological Inventories (EURPhon)

The Database of Eurasian Phonological Inventories, or EURPhon (Nikolaev 2018), which is accessible at `eurphon.info`, describes the phonological inventories of 536 Eurasian lects. It also contains some phonotactic information for many of the lects, such as word-initial consonant clusters, word-final consonants, and possible syllabic templates. This database is possibly the database that is the most similar to Phonotacticon 1.0, which also provides the phonotactic profiles of Eurasian lects, even though the two databases bear some structural differences, as will be explained in Section 3.4.

2.4.3 PHOIBLE 2.0

PHOIBLE 2.0 (Moran and McCloy 2019), which is accessible at `phoible.org`, may be the largest and the most widely used phonological database at present. Similar to UPSID but on a much larger scale, PHOIBLE 2.0 contains the phonological inventories of 2,186 lects worldwide. One of its strengths is that it often includes multiple inventories for each lect retrieved from different sources (including UPSID and EURPhon), thus, enabling cross-doculect comparisons. For instance, four inventories are available for Korean. This is quite useful considering that different descriptions of a lect's phonological inventory can vary to a significant degree depending on the consulted bibliographical source (C. Anderson et al. 2023). Unlike UPSID, it also describes the tonemes of the tonal languages.

2.4.4 PBase

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

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

As an example, for Indonesian (`pbase.phon.chass.ncsu.edu/language/4`), PBase lists /p t 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 about its phonotactic distribution, such as only /p t k ʔ s h m n ŋ l r w j/ appearing as the morpheme-final consonant. It also provides a non-exhaustive list of phonological rules, such as /p t k/ being unreleased word-finally.

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

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

2.4.5 Lyon-Albuquerque Phonological Systems Database (LAPSyD)

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

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

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

2.4.6 BDPROTO 1.1

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

2.4.7 SegBo

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

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

2.4.8 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.

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

2.4.9 Summary of phonological databases

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

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

Table 2.1: Summary of the eight databases reviewed

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

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 [uu̯]. 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 individual researchers to measure the distance in their own way.

2.5.1 Avram (1964)

Avram (1964) 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 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:

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

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) adopts her method to measure the distance between Balkan lects. His results, however, do not quite show visually perceivable areal patterns within the Balkans. This does not necessarily disqualify Postdovalova's (1966) methodology, as the Balkans is a classic example of a linguistic area and there is a high degree of similarity between Balkan lects, which could make it difficult to detect internal clusters.

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 phono-

tactics. 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 Tambovtsev (2001)

Tambovtsev (2001) compares Mongolic and Turkic lects based on the frequency of the consonant classes in each lect. When measuring the frequency of a phoneme in a lect, two types of frequency must be distinguished: *lexical frequency*, measuring the number of lexemes containing a given phoneme, and *token frequency* (also referred to as *discourse frequency* by Macklin-Cordes and Round 2020), measuring the occurrence of a given phoneme in the usage of that lect. For example, the lexical frequency of /s/ in English is the proportion of English lexemes that contain /s/ in the English vocabulary, while its token frequency is the occurrence of /s/ within the utterances of English speakers. Based on the token frequency of consonant classes, such as labials or fricatives, retrieved from the corpora of Mongolic and Turkic lects, Tambovtsev (2001) measures the phonological distance between sixteen lects.

Tambovtsev's (2001) approach of weighing the segments based on their frequency merits further exploration and may be applied to phonological characters other than consonant classes, such as vowels quality, phonotactic positions, segmental sequences, and even suprasegmental features such as tones. Since /ʒ/ occurs frequently in French but relatively infrequently in English (as well as being phonotactically limited, mostly occurring intervocalically), it can be misleading to treat the French /ʒ/ and the English /ʒ/ identically when comparing French and English phonologies. One realistic difficulty is that measuring token frequency requires corpus data, which may not be available for underdocumented lects. But when the resources permit, segmental frequency can certainly be worthy of its weight in measuring phonological distance.

2.5.5 Eden (2018)

Eden (2018) presents three types of 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, retrieved from lexical data;

- 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, most of them European. The Hamming Distances between European, Northeast Indian, and Oceanian lects show that the lects of each region are generally closer to each other phonologically. The entropy algorithm method, tested on seven European lects, shows that the similarities pattern with genealogy, Germanic lects being similar to each other and Romance lects also to each other. The spoken language identification method, tested on Greek, English, German, and Spanish, shows that Greek is closer to Spanish whereas English is closer to German. While Greek and Spanish belong to distant branches of Indo-European, English and German both belong to the West Germanic branch. All her three methodologies therefore show expected results, aligning with areality (Hamming Distance) and genealogy (enthropy algorithm and spoken language identification).

Eden 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 be used to cross-check the distances between the same lect pairs. The comparison between the results of the various methodologies (Section 7.3) shows that only some of the distances are consistent throughout all methodologies, namely the close distance between Greek and Spanish and the long distance between Germanic and Spanish/Portuguese.

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*, I 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, whereas that between /m/ and /f/ is 1/4. Thus, /f/ is the closest phoneme to /p/.

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

$$\begin{aligned} &(Jaccard(/\text{p}/, /\text{p}/) + Jaccard(/\text{f}/, /\text{p}/) + Jaccard(/\text{m}/, /\text{m}/)) \\ &+(Jaccard(/\text{p}/, /\text{p}/) + Jaccard(/\text{m}/, /\text{m}/)) \\ &= (0 + 1/4 + 0) + (0 + 0) \\ &= 1/4 \end{aligned} \tag{2.9}$$

Using this methodology, Nikolaev (2019, p. 113) shows that among the spoken lects of Eurasia, neighboring lects that are genealogically related show more similarity to each other than non-neighboring genealogically related lects and neighboring, genealogically unrelated lects also show more similarity to each other than non-neighboring, genealogically unrelated lects. He does not describe in detail which lects are similar to each other, however.

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 112 Pama-Nyungan lects "in terms of which sequences of two segments (*biphones*) they permit and which they do not" (p. 225) to measure the similarity between them and compare it to the phylogeny of those lects based on lexical cognacy. Their results show that the phonotactic information show **phylogenetic signals**, the similarity between genealogically closer lects, suggesting that phonotactic information can convey genealogical information.

The authors briefly mention, however, that one of the limits of their study is that areal-

ity was not considered as a potential factor motivating phonotactic similarity (p. 247). This may be a concerning issue, considering 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 considerably 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. Geographically speaking, this is the expected result, as Evenki is spoken mostly in Siberia and the region where (Halh) Mongolian is spoken borders Siberia, whereas Uyghur is spoken mostly in western China.

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 516 Eurasian lects. Section 3.2 explains how I chose the 516 sample lects. Section 3.3 lays out the *profile* of each lect coded in the database. Section 3.4 explains how this database is different from an existing database, EURPhon (Nikolaev 2018). Section 3.5 concludes the chapter and previews how the database will be used for the remaining part of the thesis.

3.2 Lect sampling

The 516 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); and
- which had at least one appropriate source accessible to me.

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

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

not by linguistic genealogy or cultural history. Unlike the traditional continental division between Eurasia and Oceania, the distinction is made between Eurasia, Papunesia, and Australia. "Papunesia" refers to the insular Southeast Asia plus Oceania minus Australia. Most of the southern Pacific islands, such as Taiwan, Borneo, or the Philippines, are classified as Papunesia and not Eurasia. Hainan, on the other hand, is classified as Eurasia, as it is separated only by a very thin strait from continental China. Some islands that are too small to be reflected in the resolution of Hammarström and Donohue's study are interpreted as part of a bigger landmass. For example, Ryukyu islands were too small to be reflected in the resolution and were grouped together as the Japanese archipelago, even though some Ryukyu islands are very close to Taiwan.

The distribution of the 516 sample lects is visualized in Figure 3.1, where each color-shape combination represents a family.

3.3 Phonological profile

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

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

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

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

Table 3.1: Phonological profile of A'ou

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

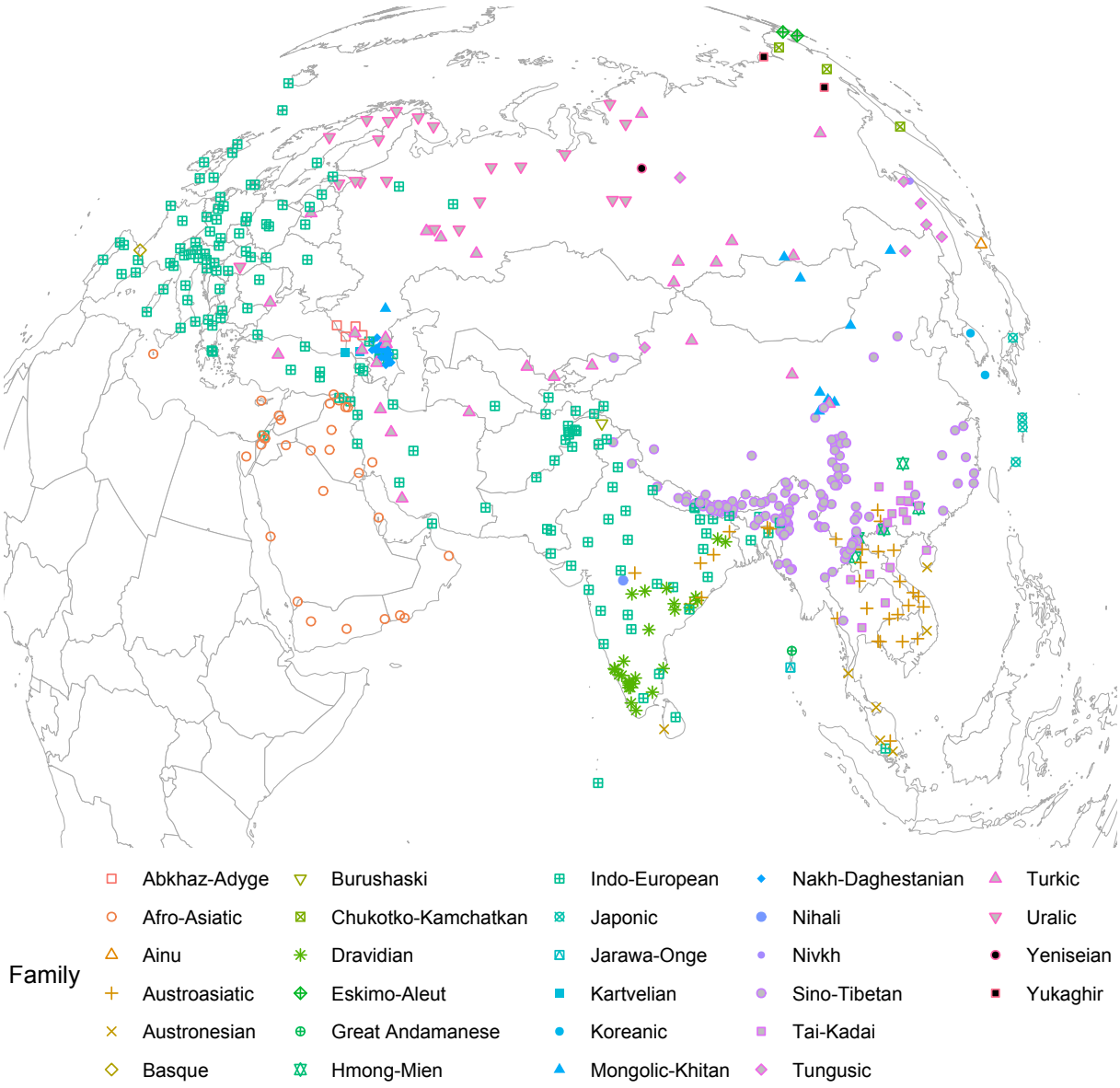


Figure 3.1: 516 Sample lects of Phonotacticon

The remaining three variables, onset, nucleus, and coda forms, were selected because they form the building units of a **syllable**, which is a concept employed by the majority of the phonological analyses of different lects (Hulst and Ritter 1999; Goldsmith 2011, cf). Malaia and Wilbur (2020) argue that the syllable is a universal strategy to divide continuous linguistic information into discrete segments, observable in spoken and the sign modalities alike. As the syllable is a widely accepted theoretical notion whose reality is supported by neurolinguistic evidence, it is the most suitable framework to be adopted for a cross-theoretical database like Phonotacticon.

An alternative to the syllable would be the notion of **word**, as many phonological analyses describe a lect's phonotactic patterns based on word boundaries, such as word-initial or word-final consonant clusters, rather than syllable boundaries. But previous works on wordhood do

not agree on what a phonological or prosodic word is, and many suggest that it is not a cross-linguistically consistent concept (Dixon and Aikhenvald 2003; Schiering et al. 2010). Thus, it is more cross-theoretically consistent to unify the variables of the database into syllabic notions rather than wordhood notions.

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 bilabial plosive (which can be aspirated or unaspirated based on its environment), using the symbol <p> overspecifies this phoneme in terms of aspiration, as English /p/ can be either aspirated (as in *pan* [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>. If there is a form in isolation, then that form is chosen as the representing segment. An example is Japanese moraic nasal /N/, which may occur as [nː], [mː], [ɴː], or others depending on phonotactic context (Iwasaki 2013). /N/ occurs as [ɴː] when it does not precede any segment (e.g. *san* さん [sɑ̃ɴː] 'three'), so I have transcribed it as /ɴː/.

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

Another problem to be addressed is the xenophone, a phoneme that only occurs in loanwords. The main complication is that the status of a xenophone can vary from fully nativized

to extremely marginal and it can sometimes be difficult to judge whether a xenophone is truly a part of a given lect's phonology. As I have mentioned in Section 2.2, some xenophones are indistinctively part of a lect's phonology, such as German /ʒ/, while some xenophones are distinctively foreign, such as German nasal vowels, which remain unstable and are often replaced by native phonemes (Wiese 2000, p. 12). Thus, whether a xenophone forms a part of the phonology of a lect is essentially a grey area, leaving me with the question of which xenophones to record in the database and which ones to leave out.

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

Phonemes that are used by only a portion of the whole speaker population of a given lect 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 /e͡ɪ/ 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 diphthongs in a given lect constitute individual phonemes or are combinations of two vowel phonemes is a matter of debate (Pike 1947; Berg 1986; Eliasson 2022) and is thus highly subject to theoretical bias. By listing all diphthongs as combinations of monophthong phonemes, I can make all the vowel phonemes compatible with PanPhon and allow cross-linguistic analysis, albeit at the sacrifice of favoring one theoretical approach to diphthongs over another. Moreover, regardless of the phonemic status of diphthongs and triphthongs, they are still listed in the nucleus part of the database, so there is no sacrifice at the descriptive level.

Exceptionally, 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 (<ⁿb>, <ⁿd>), whereas preaspirated consonants are transcribed as <h> followed by a tie bar and a segment (<h͡p>, <h͡t>).
- I have included the **fortis** (or **tense**) counterpart of all consonants, transcribed by the segmented followed by a small plus sign, as this concept is employed in works on Korean (Lee 2021), Swiss German (Fleischer and Schmid 2006), or other lects but not present in PanPhon. The feature of each fortis consonant is identical to its non-fortis counterpart, except that its [tense] feature is 1 and not 0.
- Some segments that I judge to be missing as accidental gaps were added. For example, /t͡sʼʷː/ was absent in PanPhon, even though /t͡sʼː/ and /t͡sʼʷ/ were present. As such cases are clearly gaps created by mistake, I added such segments in with appropriate feature values.

The revised version of PanPhon is available at `doi.org/10.5281/zenodo.10623743`.

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, ɯɥ> means any segment including any one of these graphemes in its IPA symbol. !<h, ɦ> means any segment not having these graphemes in its IPA symbol. Other than V, which stands for vowels, all the capital letters represent consonants or glides: N refers to nasal consonants and glides only, excluding nasalized vowels.

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

Table 3.2: The underspecified segments

Many grammars published in China that describe monosyllabic lects do not describe the lect's phonemic inventory in terms of segmental phonemes but rather in terms of *initials* (*shengmu* 聲母) and *finals* (*yunmu* 韵母), which correspond to onsets and rhymes. When consulting such grammars, 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/.

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 initial clusters or final 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 an initial cluster and a final cluster, or if a medial cluster does not contain sequences that appear as onset or coda, and if multiple possible interpretations reflect the sonority sequencing principle, then I resort to

the maximal onset principle (Kahn 1976), favoring complex onsets over complex codas. For example, if /pl/ is an initial cluster and /lp/ is a final cluster, /lpl/ is interpreted as /l.pl/, instead of /lp.l/.

- For triconsonantal or longer medial clusters, 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.

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/ *followed by any one of* /l r w j/.

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.

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 [t̚] is phonetically closer to [t] than it is to [s], I have listed /t/ as a possible Korean coda but not /s/.

3.3.2.2 Other rules on segmental transcription

- Dental consonants are transcribed with the dental diacritic (e.g. /t̪ d̪/) only when it is minimally contrastive with alveolar correspondents. Otherwise they are transcribed

without the dental diacritic (e.g. /t d/).

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

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 Bibliographical sources

The database includes the bibliographical information of the source consulted for each lect's profile. The sources are either the "long grammars" as defined by Glottolog 4.4 or any other source I deem relevant and accessible. The accessibility issue includes language barrier as well. In most cases, including when the sources were written in French, German, Japanese, or Chinese, this was not a concern, as I could read those lects. In some cases when I could not read very well the lect a source was written in, such as Russian or Finnish, I read it with the aid of machine translation.

3.3.5 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 Difference from EURPhon

Although Phonotacticon 1.0 is similar to EURPhon (Nikolaev 2018), introduced in Section 2.4.2, the two databases differ in several regards, namely:

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

3.5 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

So far, I have introduced how Phonotacticon 1.0 has been developed. Considering that it is the first database containing the possible onset, nucleus, and coda forms of a sizeable number of lects, I would like to present the possibilities this database can bring. In the following sections, I introduce some visualizations generated from Phonotacticon and discuss areal patterns observable from them. The visualized parameters are some of the most basic quantifiable information retrievable from the database, such as the length of segments within a syllable or the number of tones, and thus worthy to be briefly introduced here.

4.1 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**it/ *street*, /**spl**æʃ/ *splash*), three vowels in its nucleus position (/f**aɪə**/ *fire*, /**aʊə**/ *hour*), and four consonants in its coda position (/te**ksts**/ *texts*, /glɪ**mpst**/ *glimpsed*) (Gut 2009). English, and European lects in general, allow longer onsets, nuclei, and codas compared to other lects in the world. Hokkaido Ainu, for instance, allows only one segment in each of the three positions, its maximal syllable being CVC (Tamura 2000, p. 21).

To my knowledge, Maddieson (2013) 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 where the onset CC is stop + glide or stop + liquid), and *complex* (onsets can be clusters other than stop + glide/liquid and codas can be complex). He reports that ca. 56.% of the sample lects have a moderately complex syllable structure, ca. 30.9% have a complex syllable structure, and ca. 12.5% have a simple syllable structure. His data 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

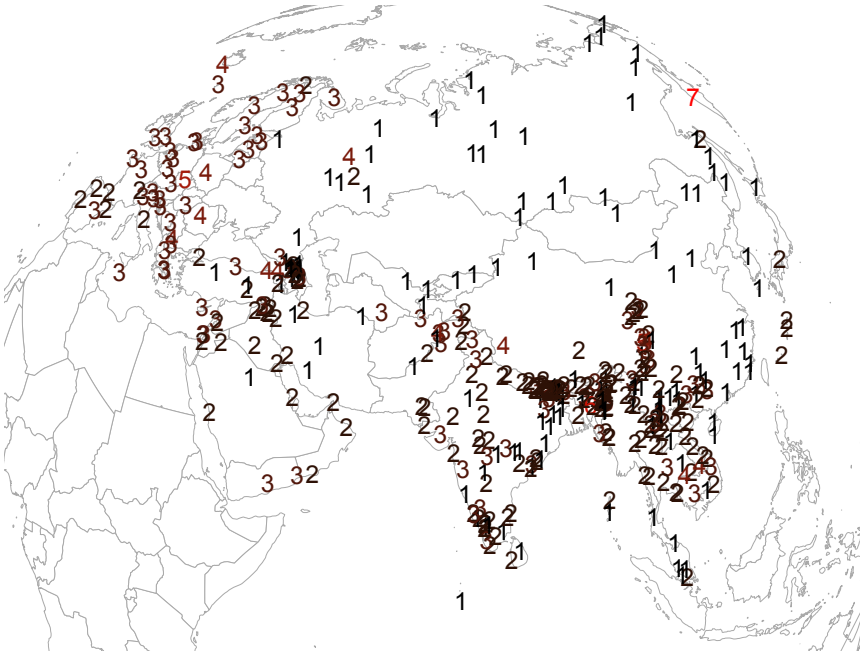


Figure 4.1: Maximal length of an onset in each lect

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 linguistic Area (J. Janhunen 2006; Dwyer 2013; Xu 2017; C. Zhou 2020, cf); South and Southeast Asia generally permit up to bisegmental onsets; and Europe generally permit up to triconsonantal onsets. The Middle East seems to be the most diverse without a dominant upper limit.

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 (Enfield 2018; Vittrant and Watkins 2019; Sidwell and Jenny 2021b, cf) and the Middle East. All sample lects that mandate an onset in a syllable use the glottal stop [ʔ] as the filler segment to fill in the gap of a syllable that would otherwise lack an onset. [ʔ] may or may not be a phoneme in such lects.

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

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,

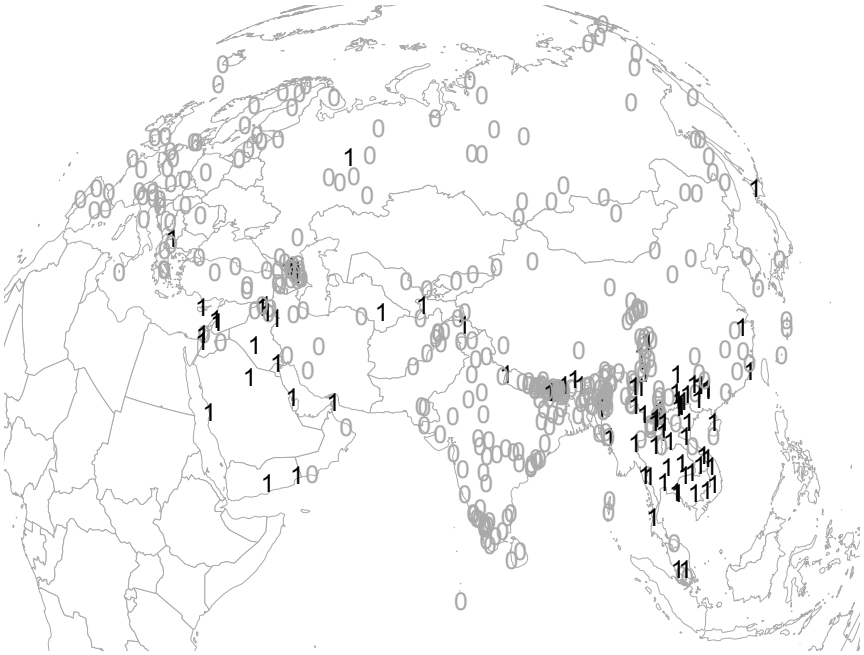


Figure 4.2: Minimal length of an onset in each lect

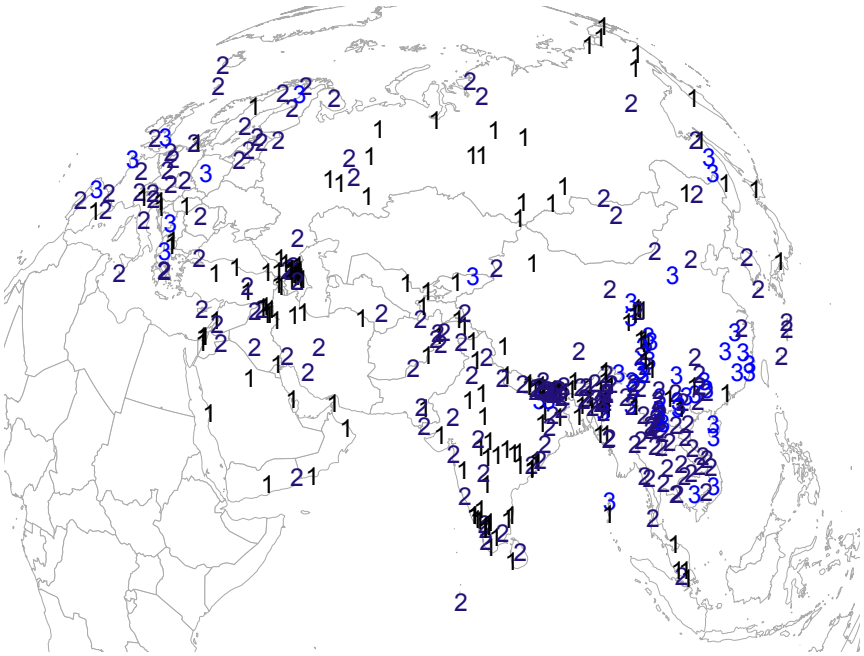


Figure 4.3: Maximal length of a nucleus in each lect

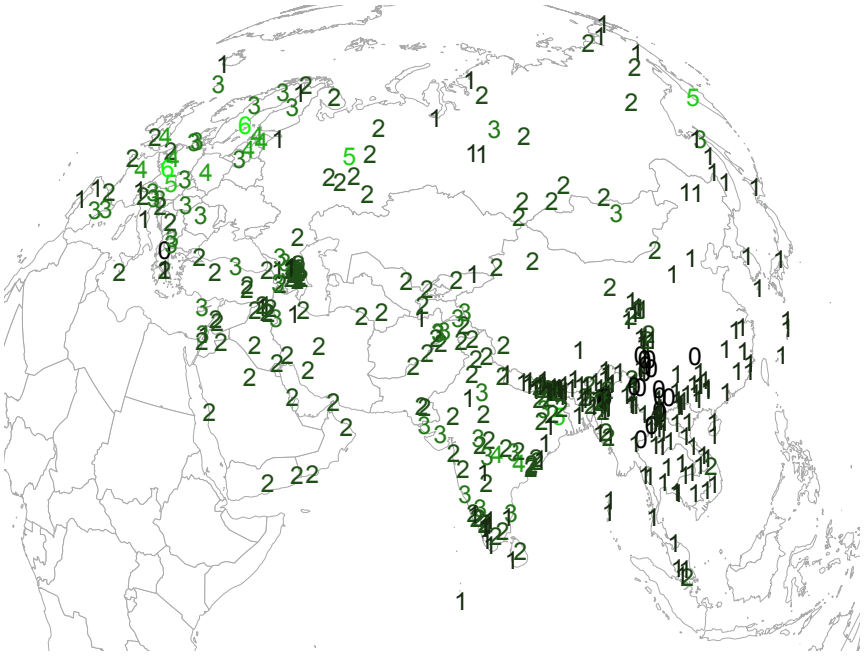


Figure 4.4: Maximal length of a coda in each lect

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

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, which both tend to increase westwards.

To confirm the visual observation that the maximal lengths of onset and coda tend to be longer in western Eurasia compared to eastern Eurasia, I have tested whether the maximal onset and coda lengths are correlated with the longitude of the Eurasian lects. First, it is necessary to test the spatial autocorrelation, as geographically neighboring lects may have similar phonotactic patterns. I identified the geographical neighbors of each lect, defined by

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

Table 4.1: Moran's I

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

Table 4.2: Spatial regression of longitude and onset/coda length

lects whose coordinates are within 1,500km distance. This distance threshold leaves no sample lect without any neighbor. I then created a weight matrix and assign the value of 1 to each neighboring lect pairs and the value of 0 to each non-neighboring lect pairs. Based on this weight matrix, I performed the Moran's I test (Table 4.1) to test the spatial autocorrelation, which confirms that both onset length and coda length are areally clustered ($p < 0.001$). Finally, based on the spatial lag model, I performed spatial regression to test the correlation between longitude of the lects and their onset/coda length. The results (Table 4.2) show that both onset and coda lengths are correlated with longitude. This confirms the visual observation that the maximal onset length and the maximal coda length grow as one goes westwards in Eurasia.

Other than geographical coordinates, it is worthwhile to compare the maximal length of onset/nucleus/coda based on language families. Figure 4.5 shows the average maximal length of onset, nucleus, and coda per each family. We see that generally, language families in western Eurasia, such as Indo-European and Afro-Asiatic, allow more segments per syllable than language families in eastern Eurasia, such as Tungusic and Sino-Tibetan.

4.2 Syllabic consonants

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

Figure 4.6 shows the distribution of lects that allow a syllabic consonant as its nucleus (blue circles) and those that do not (red crosses). We observe that syllabic consonants are generally permitted at the two extremes of Eurasia: In East and Southeast Asia and (to a much lesser degree) in Europe. Although not shown in the visualization, the phonotactic patterns of the syllabic consonants in these two areas also tend to differ. In East and Southeast Asia, syllabic nasals tend to occur as monosegmental syllables, such as Yue Chinese *m*^{4} 唔 [m̩^{21}] 'not', and syllabic fricatives tend to occur only after homoorganic fricatives, such as Mandarin

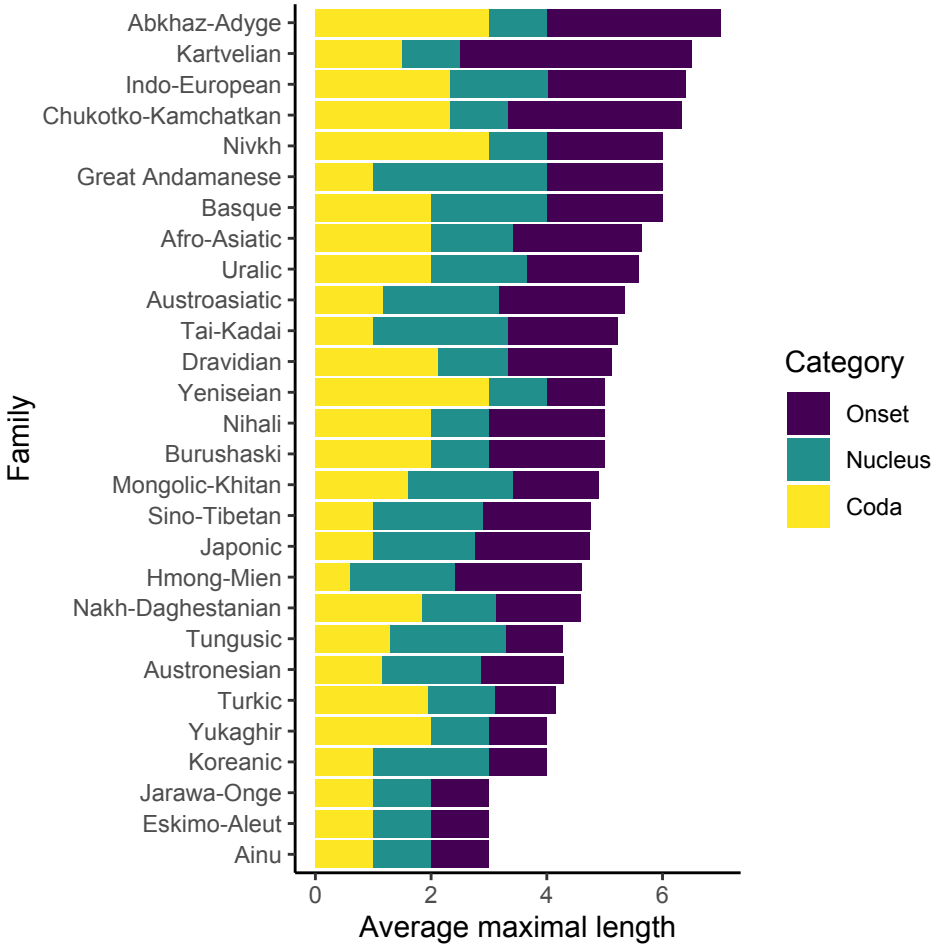


Figure 4.5: Average maximal length of onset by family

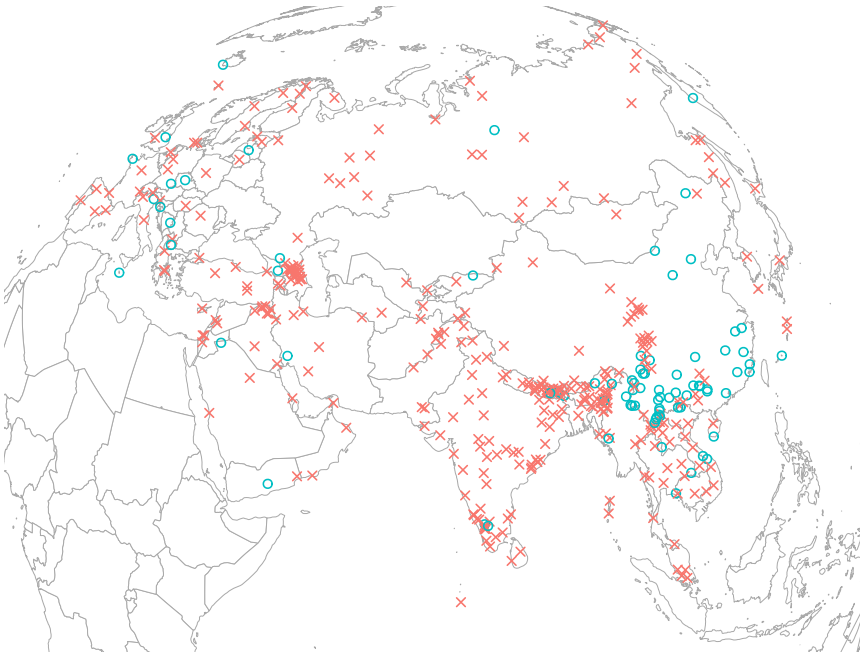


Figure 4.6: Syllabic consonants

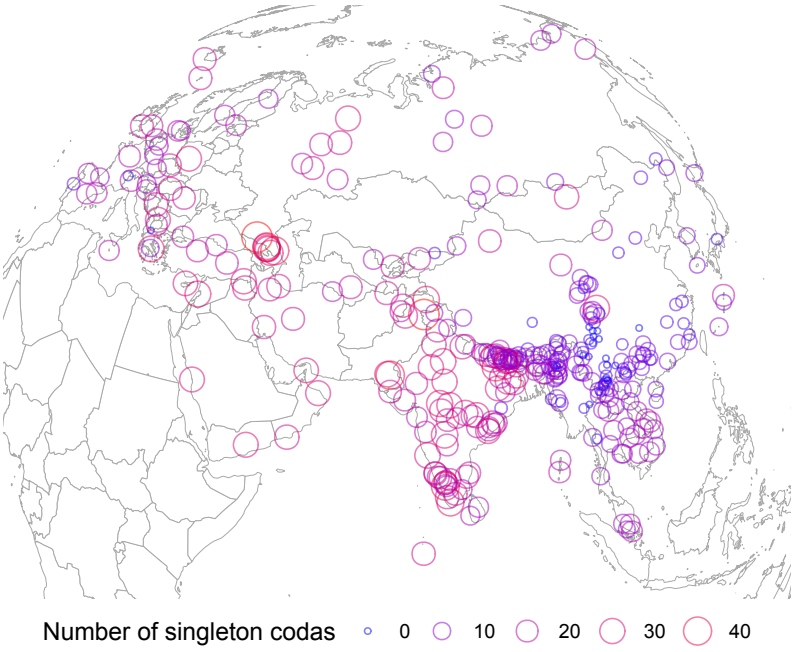


Figure 4.7: Number of singleton codas

Chinese *sì* 四 [sz̩^{51}] 'four'. In European lects, however, syllabic consonants have relatively less phonotactic restriction and can occur after a wider range of onsets, such as English *button* [bʌ.ʔn̩] or German *Vogel* [fo.gl̩] 'bird'.

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

4.3 Number of singleton codas

In Section 4.1, we saw that the maximal coda length varies across Eurasia. Codas are often limited not only quantitatively, but also qualitatively, as many lects only allow a subset of their phonemes to appear in the coda position. Although many lects also ban certain phonemes from the onset position as well, restriction in the coda position tends to be much stronger. For example, Mandarin Chinese only allows /n ŋ/ as codas, while allowing all consonant phonemes but /z ŋ/ as onsets.

Figure 4.7 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.

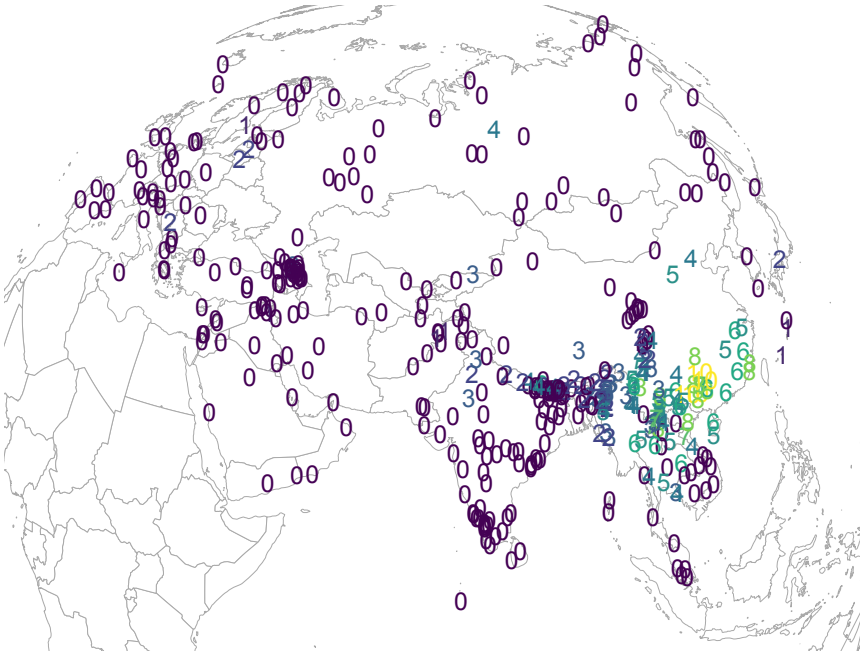


Figure 4.8: The number of tones per lect

4.4 Number of tones

Maddieson's (2013c) survey of 526 lects worldwide reveals that 220 of them are tonal. Among these tonal lects, 132 have a "simple tone system" with only two tones. The remaining 88 have a "complex tone system" with three or more tones. His data 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 4.8 shows the number of tones per Eurasian lect. The largest number of distinctive tonemes is ten, e.g. in Cao Miao (M. Wu 2015), and the lowest number is one, e.g. in Swedish, where the tonal distinction is privative, i.e. between the lexical tone and its absence (Riad 2013). It is easily observable that tones are a strongly areal phenomenon, concurring with Maddieson (2013). Most tonal lects are distributed in Mainland Southeast Asia and China (with the notable exceptions of the Qinghai-Gansu linguistic area, Cambodia, and southern Vietnam). Within this area, the Guangxi province has the highest number of tones, the maximal number being ten. Elsewhere, tones are only sparsely present, with at most two tones. From this uneven distribution, we can know that tonogenesis (the emergence of tones) is highly prone to areal pressure, even though it can happen in non-tonal environments (e.g. in Swedish).

It is worth noting that Korean, while depicted as atonal on Figure 4.8, 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.

4.5 Summary

In this section, I have shown how a number of phonological patterns vary across Eurasia. Crucially, different phonological patterns show different areal distributions: The distribution of tones (§4.4), for example, is not identical to the distribution of syllabic consonants (§4.2). It is therefore helpful to shed light on each one of the phonological patterns to understand their diverse areal shapes.

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 5.2 shows the whole process of measuring structural phonological distance. Section 5.3 uses Grambank (Skirgård et al. 2023), a morphosyntactic database to calculate the morphosyntactic distances between Eurasian lects and compare them to the phonological distances. Section 5.4 compares the phonological distances and morphosyntactic distances to genealogical relatedness, defined as the number of shared genealogical layers between two lects of the same family. Section 5.5 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. It is somewhat similar to Nikolaev's (2019) methodology (§2.5.6) in that it measures the distance between two sets of sounds based on the distances between each sound in one set and its featurally closest sound in the other set. The crucial difference lies in that Nikolaev's (2019) distance measure is between phonemic inventories, while the following analysis measures the distance between onset, nucleus, and coda sequences, as well as the numbers of tones. This allows us to calculate the distance between different phonologies based not only on what phonemes a lect has, but also on what phonemes can appear where within a syllable in relation to other phonemes. For example, a inventory-based method cannot capture the difference between a lect having /ʔ/ only as an onset, a lect having it only as a coda, and a lect having it in both positions. The syllable-based method employed here can help overcome that limitation.

5.2.1 The data

The data consists of each lect's phonological *sequences* in the onset, nucleus, or coda position, which are one or more phonological *segments* that belong to the phonemic inventory of the lect. For example, /spl/ is an onset sequence of English, composed of three segments: /s/, /p/, and /l/.

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 onverting [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 [pbfvmtdszncɟʃʒɲkgxhlrr̝j][pbtdgfvszʃʒxjrr̝lmnɲ]. 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.

5.2.2 Measuring the distance between sequence

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

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
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
Sum			5

Table 5.1: The Saporta distance between /t/ and /p/

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, Table 5.1 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/.

Although the Saporta distance is the distance between two segments and not sequences, I will apply it to measure the distance between sequences. For example, in order to compare the distance between /pl/ and /spl/, I calculate the distance between all the logically possible mappings between the two sequences, as shown in Table 5.2.

The distance between each pair of segments mapped onto the same position is compared. An empty slot is considered to be a segment that has the 0 value for all phonological features. Among the possible mappings, the third one yields the minimal distance between /spl/ and /pl/, as /p/ is compared to /p/, /l/ is compared to /l/, and /s/ is compared to a non-existing

Mapping 1		s	p	l	
	p	l			
Mapping 2	s	p	l		
	p	l			
Mapping 3	s	p	l		
		p	l		
Mapping 4	s	p	l		
			p	l	
Mapping 5	s	p	l		
				p	l

Table 5.2: Five possible mappings between /spl/ and /pl/

segment with 0 values for all features. Thus, the second comparison is chosen as the distance between /spl/ and /pl/. 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 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.[1]

As an example, Table 5.3a shows the twenty sequences that are the most similar to /pl/ and Table 5.3b those to /ia/. Note that [a], [ă], and [æ] are not featurally distinct in PanPhon, as they are all low front unrounded vowels. The distance between /ia/, /iă/, and /iæ/ is thus zero.

5.2.3 Measuring the segmental distance between lects

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

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 5.4. 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 (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.

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

Sequence	Distance
pl	0
pˣl	1
bl	2
pʰl	2
p̥l	2
pˀl	2
ⁿpl	2
pˀl	2
plʲ	2
fl	3
p͡fl	3
plˠ	4
m̥l	4
b̥l	4
b̤l	4
pʰl̥	4
pz	4
ⁿbl	4
bˀl	4
ɓl	4

(a) /pl/

Sequence	Distance
ia	0
iă	0
iæ	0
ie	2
ea	2
iã	2
iĕ	2
iæ̃	2
ɪa	2
iaː	2
iːa	2
ĩa	2
ɨa	2
ią	2
eæ	2
iæː	2
iːæ	2
iɑ	3
iɛ	4
iːe	4

(b) /ia/

Table 5.3: Sequences the most similar to /pl/ and /ia/

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

Table 5.4: The comparison between Lect A with the onset sequences /p m t/ and Lect B with the onset sequences /p t/

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

5.2.4 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:

$$\frac{|T1 - T2|}{T1 + T2} \tag{5.1}$$

For example, Burmese has 3 tonemes, whereas Yue Chinese has 6. The tonal distance between two lects is thus:

$$\frac{|3 - 6|}{3 + 6} = \frac{1}{3} \tag{5.2}$$

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

5.2.5 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).

Admittedly, assigning equal weight to the four types of distances is a rather simplistic approach. As a reviewer of this dissertation pointed out, given the wider articulatory variance of consonants compared to that of vowels, the distance of onsets and codas, which mainly consist of consonants, may merit more weight than the distance of nuclei, which mainly consist of vowels. Additionally, weighing the tonal distance equally to the three segmental distances may also be a problem, since unlike the segmental distances, the tonal distances are not normally distributed: as there are more atonal lects than tonal lects, the tonal distances are unevenly distributed to the two extremes of 0 (two atonal lects or two tonal lects with the same number of tones) and of 1 (a tonal lect versus an atonal lect). A more sophisticated weighing of the four types of distances that does justice to the different levels of variance and distribution of the four parameters is warranted in the future analyses of phonological distance using Phonotacticon.

Table 5.5 shows the twenty lect pairs with the smallest distance value. We see that, unsurprisingly, some of the closest lects are those that are spoken by the same ethnic group or closely related groups, namely Yukaghir (Northern and Southern), Korean (Seoul and Jeju), and Lao-Isan (Lao and Northeastern Thai).

While I cannot explain the distance between every lect pair in detail, I can focus on several lects and the lects they are the most similar to. In the following sections, I will discuss ten lects and the lects that are phonologically similar to each of them. The ten lects were selected based

Lect vs. Lect	Distance
Atong (India) vs. Standard Malay	0.26
Northern Yukaghir vs. Southern Yukaghir	0.26
Kelantan-Pattani Malay vs. Tundra Nenets	0.29
Jejueo vs. Korean	0.32
Lao vs. Northeastern Thai	0.32
Chitwania Tharu vs. Wambule	0.34
Biyo vs. Kaduo	0.34
Kelantan-Pattani Malay vs. Standard Malay	0.36
Lak vs. North-Central Dargwa	0.37
North-Central Dargwa vs. Rutul	0.38

Table 5.5: The ten lect pairs with the shortest phonological distance

on what I assume the readers may find theoretically interesting and may provide insight on the typological history of Eurasia and language contact in general.

(Continued on the next page)

5.2.5.1 Basque

Basque is a lect isolate spoken in northern Spain and southern France, spoken by the Basque minority and surrounded by the dominant Indo-European lects, namely French and Spanish. Given its geographical and sociolinguistic position, one would assume it to be close to the European lects, especially the Romance lects surrounding it.

Family	Lect	Distance
Indo-European	Daman-Diu Portuguese	1.27
Indo-European	Ladino	1.33
Sino-Tibetan	Sunwar	1.35
Indo-European	Catalan	1.45
Indo-European	Assamese	1.55
Austroasiatic	Gata'	1.57
Sino-Tibetan	Bunan	1.65
Indo-European	Welsh	1.68
Indo-European	Nimadi	1.68
Indo-European	Dutch	1.70
Indo-European	Estonian Swedish	1.70
Indo-European	Maithili	1.80
Indo-European	Arvanitika Albanian	1.80
Indo-European	Italian	1.83
Indo-European	Kashmiri	1.84
Uralic	Pite Saami	1.86
Sino-Tibetan	Thangmi	1.87
Indo-European	Northern Pashto	1.89
Sino-Tibetan	Bujhyal	1.92
Indo-European	Godwari	1.96

Table 5.6: Twenty lects closest to Basque

Table 5.6 shows the twenty Eurasian lects phonologically most similar to Basque. Eight of them are European, not including one European-based creole spoken in India (Daman-Diu Portuguese). The convergence between Basque and the neighboring (Western) Romance lects has been well discussed (Jendraschek 2019). The results show, however, that the convergence between Basque and European lects is not limited to Romance lects but also to the lects of other Indo-European branches, such as Celtic or Germanic, or non-Indo-European European, such as Pite Saami. This does not necessarily imply direct contact between Basque and these European lects, but rather that they belong to the same European linguistic area.

5.2.5.2 Evenki

Evenki is a Tungusic lect spoken by the Evenks throughout the vast region encompassing Siberia, Mongolia, and northeast China. Anderson (2006), who argues for a Siberian linguistic area, suggests that Evenki may have been the carrier of some of the Siberian areal features, given the high mobility of its speakers. Evenki itself also shows influence from Mongolic (Poppe 1972; Khabtagaeva 2010) and Turkic (namely Sakha; Pakendorf 2020).

Family	Lect	Distance
Turkic	Kirghiz	0.51
Dravidian	Konda-Dora	0.59
Turkic	Tatar	0.59
Yukaghir	Northern Yukaghir	0.63
Turkic	Tuvinian	0.67
Indo-European	Sadri	0.68
Turkic	Kazakh	0.69
Yukaghir	Southern Yukaghir	0.71
Turkic	Uighur	0.75
Nakh-Daghestanian	Budukh	0.75
Dravidian	Ravula	0.79
Uralic	Komi-Zyrian	0.84
Indo-European	Halbi	0.84
Turkic	Chuvash	0.85
Turkic	Southern Altai	0.89
Indo-European	Kotia-Adivasi Oriya-Desiya	0.97
Jarawa-Onge	Jarawa (India)	0.99
Turkic	Sakha	1.04
Burushaski	Burushaski	1.04
Turkic	South Azerbaijani	1.04

Table 5.7: Twenty lects closest to Evenki

Table 5.7 shows the twenty lects closest to Evenki. As expected, the closest lects are mostly spoken in northern Eurasia, namely Turkic, Uralic, and Yukaghir lects. Surprisingly, none of the top twenty are other Tungusic sample lects (Manchu, Nanai, Negidal, Oroch, and Udihe). This suggests that Evenki might have diverged from the rest of the Tungusic family in favor of converging into other northern Eurasian lects, especially Turkic.

5.2.5.3 Georgian

Georgian is the largest lect of the Kartvelian family and the official lect of Georgia. Caucasus, the region where Georgia is located, is a hotspot of linguistic diversity. Language families local to Caucasus include not only Kartvelian, but also Nakh-Dagestanian, Abakhaz-Adyghe, Indo-European, or Turkic.

Family	Lect	Distance
Sino-Tibetan	Japhug	3.32
Indo-European	Macedonian	3.41
Kartvelian	Laz	3.53
Austroasiatic	Laven	3.85
Indo-European	Icelandic	3.91
Sino-Tibetan	Situ	3.95
Sino-Tibetan	Chak	3.96
Sino-Tibetan	Zbu	4.03
Indo-European	Gheg Albanian	4.09
Indo-European	Russian	4.16
Indo-European	Piemontese	4.17
Indo-European	Dutch	4.18
Dravidian	Tulu	4.19
Sino-Tibetan	Purik-Sham-Nubra	4.25
Indo-European	Nuristani Kalasha	4.28
Indo-European	Arvanitika Albanian	4.29
Abkhaz-Adyge	Kabardian	4.30
Indo-European	Friulian	4.31
Austroasiatic	Khasi	4.32
Afro-Asiatic	Cypriot Arabic	4.32

Table 5.8: Twenty lects closest to Georgian

Among the twenty lects phonologically closest to Georgian shown in Table 5.8, only two is spoken in Caucasus: Laz, which also belongs to the Kartvelian family, and Kabardian, which belongs to the Abkhaz-Adyge family. No Nakh-Daghestanian lect can be found among the top twenty, even though there are thirteen Nakh-Daghestanian lects in the sample. Turkic lects spoken in Caucasus, such as Azerbaijani, are also absent. The results suggest that the phonological convergence in Caucasus is moderate as best.

5.2.5.4 Hindi

Hindi is an Indo-Aryan lect spoken mainly in the northern part of South Asia. Urdu, a similar lect spoken mainly in Pakistan, is mutually intelligible with Hindi. Urdu is not present in Phonotacticon 1.0, however. Given the previous theories on South Asia as a linguistic area (reviewed in Section 2.3.2.6), it is likely that Hindi shows a high degree of similarity with other lects of the Indian subcontinent.

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

Table 5.9: Twenty lects closest to Hindi

Table 5.9 shows twenty lects closest to Hindi. Unsurprisingly, the majority of the lects are spoken in South Asia. They include not only sister lects belonging to the (Indo-Aryan branch of) Indo-European family, but also the other lects spoken across South Asia, namely Dravidian, Austroasiatic, Sino-Tibetan, and the lect isolates Nihali and Burushaski. This clearly demonstrates the high degree of convergence between Hindi and other lesser-spoken lects of India, although the directionality of convergence is less clear.

5.2.5.5 Japanese

Japanese is the most widely spoken Japonic lect, whose prestige form is based on the Tokyo variety. While it is the dominant lect in the Japanese peninsula, the southern and northern edges of the peninsula are home to Ryukyuan and Ainuic lects, respectively. Thus, we may expect that Japanese shows some degree of phonological similarity to other lects spoken in Japan and neighboring areas, such as Korea, Northeast China, and Russian Far East.

Family	Lect	Distance
Sino-Tibetan	Nocte Naga	1.17
Sino-Tibetan	Darma	1.24
Sino-Tibetan	Lamjung-Melamchi Yolmo	1.26
Sino-Tibetan	Kado	1.33
Sino-Tibetan	Thakali	1.36
Sino-Tibetan	Maram Naga	1.37
Sino-Tibetan	Moyon	1.41
Sino-Tibetan	Kyerung	1.52
Sino-Tibetan	Sangkong	1.56
Austroasiatic	Chong of Chanthaburi	1.57
Sino-Tibetan	Chothe	1.59
Sino-Tibetan	Pwo Eastern Karen	1.62
Sino-Tibetan	Deori	1.63
Sino-Tibetan	Galo	1.65
Sino-Tibetan	Tibetan	1.66
Sino-Tibetan	Zaiwa	1.68
Sino-Tibetan	Southern Jinghpaw	1.69
Sino-Tibetan	Asho Chin	1.70
Sino-Tibetan	Khams Tibetan	1.71
Sino-Tibetan	Western Parbate Kham	1.73

Table 5.10: Twenty lects closest to Japanese

Strikingly, however, we see from Table 5.10 that none of the twenty lects that are phonologically closest to Japanese neighbors Japanese or even belongs to the same Japonic family. Instead, all the closest lects belong to the Tibeto-Burman branch of the Sino-Tibetan family, except for Chong of Chanthaburi (Austroasiatic). From this non-areal pattern, we can conclude that Japanese phonology is an "odd one out" that does not bear much similarity to its neighbors and that its phonotactic patterns are strongly Tibeto-Burman.

5.2.5.6 Kazakh

Kazakh is a Turkic lect spoken in Central Asia. The Turkic family has been known to share a wide range of similarities with neighboring families, namely Mongolic and Tungusic, such that the three families have been often grouped together as the Altaic family, which remains controversial and is not accepted by most linguists today. For a recent review of Altaic, see Janhunen (2023), who sees Altaic not as a family but as a typological sphere.

Family	Lect	Distance
Turkic	Kirghiz	0.43
Turkic	Tatar	0.45
Turkic	Uighur	0.68
Tungusic	Evenki	0.69
Indo-European	Halbi	0.73
Dravidian	Konda-Dora	0.77
Uralic	Komi-Zyrian	0.78
Indo-European	Sadri	0.83
Nakh-Daghestanian	Budukh	0.89
Turkic	Chuvash	0.89
Turkic	Kumyk	1.04
Indo-European	Kotia-Adivasi Oriya-Desiya	1.05
Turkic	Tuvinian	1.06
Nakh-Daghestanian	Southwestern Dargwa	1.13
Nakh-Daghestanian	Avar	1.17
Burushaski	Burushaski	1.17
Turkic	Southern Altai	1.17
Nakh-Daghestanian	North-Central Dargwa	1.19
Nakh-Daghestanian	Rutul	1.19
Yukaghir	Northern Yukaghir	1.20

Table 5.11: Twenty lects closest to Kazakh

Table 5.11 shows the twenty lects most similar to Kazakh. Many of the twenty closest lects are other Turkic lects and also non-Turkic lects spoken in central and northern Eurasia, such as Avar, Budukh, Evenki, Rutul, Dargwa, Yukaghir, and Komi-Zyrian. It does not show as high level of similarity to Mongolic phonology, however.

The similarity of Kazakh (and perhaps Turkic in general) to other families not commonly classified as Altaic, namely Uralic, Yukaghir, and Nakh-Daghestanian, raises the question whether Altaic is useful even as a typological concept. Given that the widespread phonological convergence in central Eurasia is not limited to Turkic, Mongolic, and Tungusic, it remains doubtful whether these three families have distinct typological characteristics not also shared by its geographical neighbors. An areal classification, such as Northeast Asia, may be more appropriate. As Janhunen (2023) points out, typological features of Altaic is not limited to the families classified as Altaic (which sometimes include Japonic and Koreanic as well) and the

features shared by the Altaic families at the proto-language level are rather scant and universally common.

5.2.5.7 Mandarin Chinese

Mandarin Chinese is the standard form of Chinese based on the Beijing variety and currently the largest lect spoken in the world, with close to a billion native speakers. Given its overwhelming demographic hegemony, one would predict that its neighboring lects in Northeast Asia would have received substantial influence from Mandarin in their phonological characteristics.

Family	Lect	Distance
Sino-Tibetan	Cosao	0.91
Sino-Tibetan	Thado Chin	0.94
Sino-Tibetan	Pela	1.03
Austroasiatic	Bolyu	1.05
Sino-Tibetan	Zeme Naga	1.08
Sino-Tibetan	Phom Naga	1.10
Sino-Tibetan	Lashi	1.12
Sino-Tibetan	Sadu	1.15
Sino-Tibetan	Ao Naga	1.28
Tai-Kadai	Cao Miao	1.31
Sino-Tibetan	Biyo	1.31
Sino-Tibetan	Kucong	1.31
Sino-Tibetan	Kaduo	1.31
Tai-Kadai	Western Ong-Be	1.31
Sino-Tibetan	Koireng	1.32
Austroasiatic	Bugan	1.35
Tai-Kadai	Lao	1.38
Tai-Kadai	Duoluo Gelao	1.40
Sino-Tibetan	Honi	1.41
Sino-Tibetan	Bisu	1.43

Table 5.12: Twenty lects closest to Mandarin Chinese

Table 5.12 shows twenty lects closest to Mandarin. What is surprising is that all top twenty lects are spoken in Mainland Southeast Asia and none are Northeast Asian lects that have had contact with Mandarin throughout history, such as Tungusic, Mongolic, or Koreanic lects. Mandarin phonology is thus clearly Mainland Southeast Asian and it is phonologically an outlier in Northeast Asia, despite being the largest lect spoken there. It can be seen as an northward "extension" of Mainland Southeast Asia. While there is no doubt that Mandarin has had extensive contact with its neighboring lects, as evidenced by the extensive amount of loanwords in those lects, there seems to have been very little convergence in the phonological domain, suggesting the possibility of divergence instead.

5.2.5.8 Sri Lanka Malay

Sri Lanka Malay is a Malayic lect spoken in Sri Lanka. Malay speakers were brought to Sri Lanka by the Dutch and the British from the 17th till the 19th century (Bakker 2006). Despite the relatively recent arrival of Malay, it seems to have developed remarkably strong areal characteristics shared by other Sri Lankan lects.

Family	Lect	Distance
Indo-European	Domari	0.84
Dravidian	Tulu	0.86
Dravidian	Malavedan	0.88
Dravidian	Muduga	0.94
Indo-European	Sindhi	0.94
Austroasiatic	Pnar	1.02
Sino-Tibetan	Limbu	1.09
Dravidian	Kui (India)	1.13
Dravidian	Jennu Kurumba	1.13
Uralic	Pite Saami	1.13
Dravidian	Kodava	1.13
Indo-European	Vaagri Booli	1.16
Sino-Tibetan	Amdo Tibetan	1.17
Indo-European	Nuristani Kalasha	1.19
Dravidian	Korra Koraga	1.24
Dravidian	Sholaga	1.31
Indo-European	Konkan Marathi	1.36
Sino-Tibetan	Tangam	1.36
Sino-Tibetan	Leh Ladakhi	1.38
Dravidian	Malayalam	1.39

Table 5.13: Twenty lects closest to Sri Lanka Malay

Table 5.13 shows the twenty lects closest to Sri Lanka Malay, which does not include other Malayic lects in the database (Standard Malay, Baba Malay, and Kelantan-Pattani Malay) but is rather dominated by other lects of South Asia, including Indo-Aryan, Dravidian, and Tibeto-Burman, and Austroasiatic lects. This is clear evidence that a lect can significantly diverge from its closest sibling lects and converge into its linguistic area even within a period of time as short as two or three centuries. Bakker (2006) also considers the convergence of Sri Lanka Malay into the Sri Lankan linguistic area "a radical typological change" (p. 139), which happened not only in phonology but also in morphosyntax, namely the development of SOV word order and nominal case marking.

5.2.5.9 Standard Malay

Standard Malay, commonly simply referred to as Malay, is the form of Malay as the official lect of Malaysia. Malay is often considered to be a lect spoken at the southern end of the Mainland Southeast Asian linguistic area (Enfield and Comrie 2021; Matisoff 2019; Nomoto and Ling Soh 2019). Sidwell and Jenny (2021a), however, excludes Malay from MSEA, for it "retains much of its inherited [Austronesian] typology" (p. 3) rather than having converged into the MSEA features.

Indeed, Malay does not seem to bear most of the typical phonological features of Mainland Southeast Asia. It is not tonal; its vowel inventory is not complex; it allows most of its phonemes at the coda position; it does not have the creaky voice register; and it is not predominantly monosyllabic or sesquisyllabic. Thus, at least at the phonological level, it seems likely that Malay cannot be considered part of the Mainland Southeast Asian area.

Family	Lect	Distance
Sino-Tibetan	Atong (India)	0.26
Austronesian	Kelantan-Pattani Malay	0.36
Uralic	Tundra Nenets	0.39
Jarawa-Onge	Jarawa (India)	0.40
Chukotko-Kamchatkan	Koryak	0.41
Ainu	Hokkaido Ainu	0.42
Austroasiatic	Semelai	0.46
Chukotko-Kamchatkan	Chukchi	0.49
Mongolic-Khitan	Bonan	0.50
Uralic	Vach-Vasjugan	0.53
Turkic	South Azerbaijani	0.53
Dravidian	Ravula	0.53
Uralic	Selkup	0.59
Nakh-Daghestanian	Tsez	0.60
Turkic	Tuvinian	0.62
Mongolic-Khitan	Dagur	0.63
Eskimo-Aleut	Naukan Yupik	0.63
Tungusic	Negidal	0.63
Dravidian	Badaga	0.66
Tungusic	Udihe	0.69

Table 5.14: Twenty lects closest to Standard Malay

Table 5.14 shows the twenty lects closest to Standard Malay. Among the twenty lects, the two lects are spoken in what can be categorized as Mainland Southeast Asia: Kelantan-Pattani Malay, which is also Malayic, and Semelai, an Austroasiatic lect spoken in the Malay Peninsula. The results clearly suggests that Malay phonology is not similar to other Mainland Southeast Asian phonologies at any meaningful level. This suggests that at least in the phonological domain, the Malay Peninsula is the southern boundary of the Mainland Southeast Asia.

5.2.5.10 Tsat

Tsat is an Austronesian lect spoken in Hainan, a southern Chinese island, by the Utsul people. It belongs to the Chamic branch of the Austronesian family. Other Chamic lects are spoken by the Cham people in southern Vietnam and Cambodia, some of who immigrated to Hainan around 986 CE (Thurgood et al. 2014, Ch. 3) and later formed a distinct ethnic identity as well as a distinct variety of Chamic.

Thurgood et al. (2014, Ch. 10) describes the contact-driven influence on Tsat, mainly from Hlai (Tai-Kadai) and Southwestern Mandarin, the two dominant lects of Hainan. The influence is attested in all domains: Tsat is monosyllabic and tonal (phonology); a quarter of the vocabulary is from Mandarin, including grammatical words (lexicon); and its relative clauses can be either postnominal due to Austronesian inheritance or prenominal like Mandarin (syntax).

Family	Lect	Distance
Sino-Tibetan	Gan Chinese	0.76
Sino-Tibetan	Hakka Chinese	0.78
Sino-Tibetan	Paite Chin	0.78
Tai-Kadai	Western Ong-Be	0.80
Sino-Tibetan	Min Nan Chinese	0.90
Sino-Tibetan	Lashi	0.92
Sino-Tibetan	Hui Chinese	1.01
Austroasiatic	Bugan	1.04
Sino-Tibetan	Cosao	1.05
Sino-Tibetan	Min Bei Chinese	1.06
Sino-Tibetan	Southern Qiang	1.10
Tai-Kadai	Central Hongshuihe Zhuang	1.10
Sino-Tibetan	Jinyu Chinese	1.11
Sino-Tibetan	Northern Pinghua	1.19
Sino-Tibetan	Kucong	1.20
Austroasiatic	Mang	1.21
Tai-Kadai	Duoluo Gelao	1.24
Sino-Tibetan	Pu-Xian Chinese	1.26
Sino-Tibetan	Honi	1.26
Sino-Tibetan	Thado Chin	1.26

Table 5.15: Twenty lects closest to Tsat

Among the twenty lects closest to Tsat (Table 5.15), it is surprising that no Austronesian lect can be found, including Bih, a Chamic lect spoken in southern Vietnam. Although Hlai and Southwestern Mandarin are not represented among the sample lects, it is clear that Tsat is the most similar to the Sinitic lects: eight of the twenty closest lects are Sinitic, including northern Sinitic (Jinyu Chinese). Thus, we can conclude that Tsat phonology has strongly converged into Sinitic phonology over the past millenium.

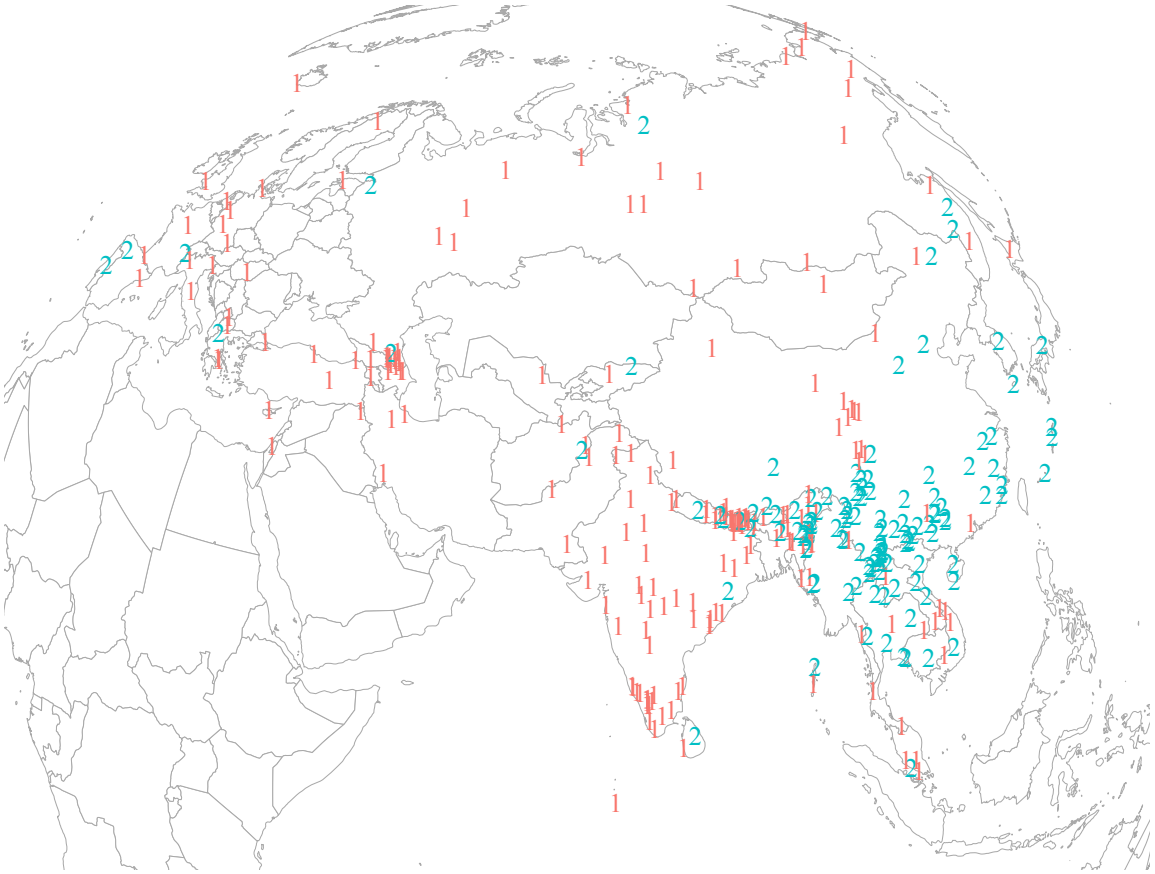


Figure 5.1: Two phonological clusters of Eurasia

5.2.6 Clustering the lects

Based on these distances, I cluster similar lects into a few groups to investigate areal patterns, using the *k*-means clustering method. The *k*-means clustering is a statistical method to divide observations into *k* number of clusters such that the variance within each cluster is minimized. Based on the the distances of each lect from other lects, I can group the lects into any number of clusters to see if those clusters are also clustered geographically on a map of Eurasia, thereby forming phonological areas. Since fewer clusters are more reliable than larger number of clusters, here I only show the first three numbers of clustering: two (Figure 5.1), three (Figure 5.2), and four (Figure 5.3).

From the visualized *k*-means clustering, we observe that phonological clusters also tend to form geographical clusters, confirming the prediction of the areality of phonological convergence among the lects of Eurasia. These clusters do not, however, always neatly fit into the previously hypothesized linguistic areas (Section 2.3).

Based on the clustering, Mainland Southeast Asia seems to go as northward as north Sinitic (Jin and Mandarin) without including Qinghai-Gansu, as I have suggested in Section 5.2.5.7 that north Sinitic is an outlier in Northeast Asia for being the northward extension of Mainland Southeast Asia. Note that, however, there seems to be a weak north-south contrast within Mainland Southeast Asia, resulting in Mainland Southeast Asia proper and southern (plus

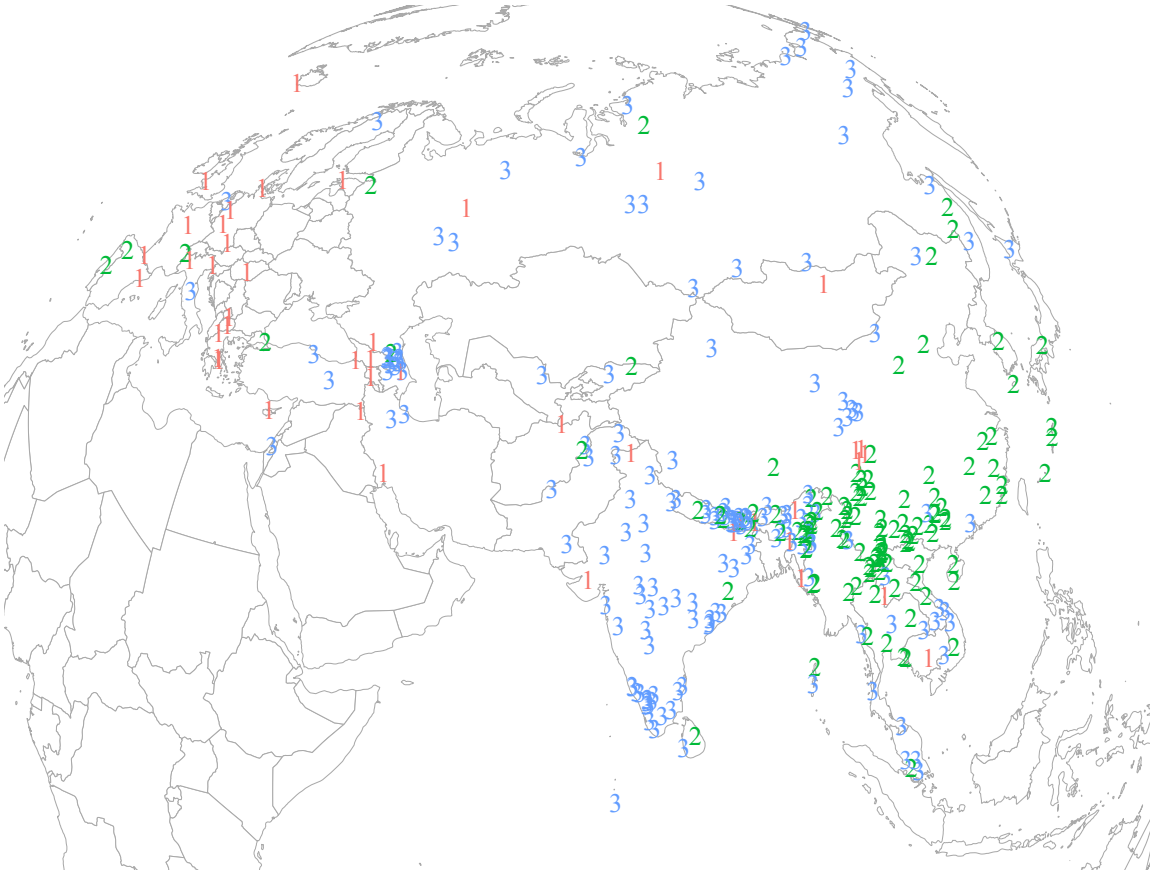


Figure 5.2: Three phonological clusters of Eurasia

some parts of northern) China, although these two regions overlap greatly. The southern limit of Mainland Southeast Asia goes as southward as central Thailand but not to the Malay Peninsula, as argued in Section 5.2.5.9 that Malay is not part of this linguistic area.

Eastwards, Mainland Southeast Asia stretches as far as the Chinese east coast and even (to some degree) the Ryukyuan islands. Eastern India and Bangladesh seems to be its western limit.

The space of Europe as a phonological area includes what is typically defined as Europe plus Anatolia. Russian seems to be the western extension of Europe as a phonological area while the non-Indo-European lects of eastern Russia are not part of it, as the level of phonological convergence between Russian and the minority lects of Russia is low.

The clustering suggests that Qinghai-Gansu is a quite distinct from lects in eastern China. South Asia seems to be not fully homogeneous, since it is divided into multiple clusters in Figure 5.3. Northeast Asia may be defined as the wide area consisting of Russia (not including Russian), parts of central Asia, Mongolia, northeast China (not including Sinitic), Korea, and Hokkaido (Ainu).

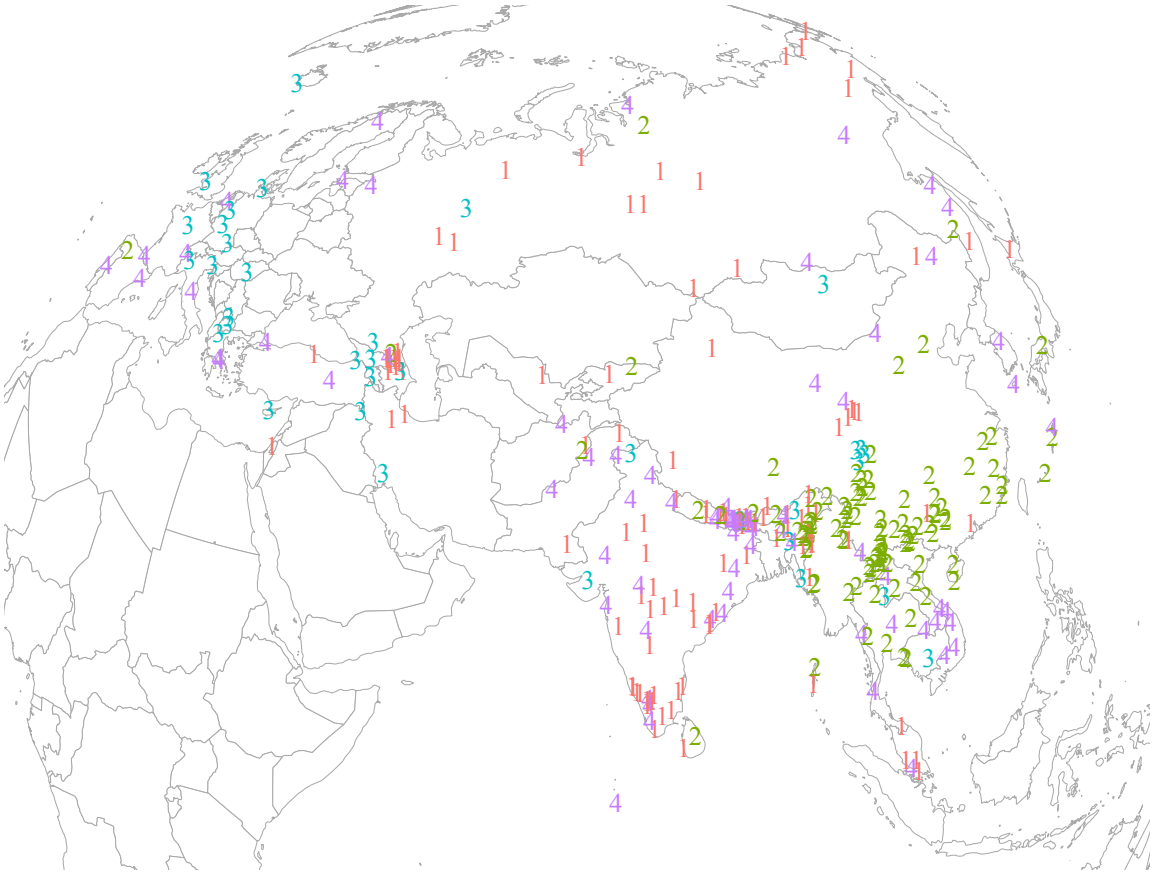


Figure 5.3: Four phonological clusters of Eurasia

5.2.7 Correlation between geographical distance and phonological distance

Next, I will test the following hypothesis: Geographical distance is correlated with phonological distance. In other words, the closer two lects are geographically, the closer they tend to be phonologically.

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

As shown in Section 5.2.6, geographically close lects also tend to show similar phonological patterns. Thus, spatial autocorrelation must be tested for the lects' distances to Burushaski. First, I define the geographical neighbors of each lect. The neighbors of each lect are defined as those whose geographical coordinates are within the maximal distance of a lect to its nearest lect. In other words, I defined neighborhood such that every lect will have at least one geographical neighbor.

I then generate a spatial weight matrix based on binary weight. The value of 1 is assigned to each neighboring lect pair, while the value of 0 is assigned to other pairs. Moran's I test

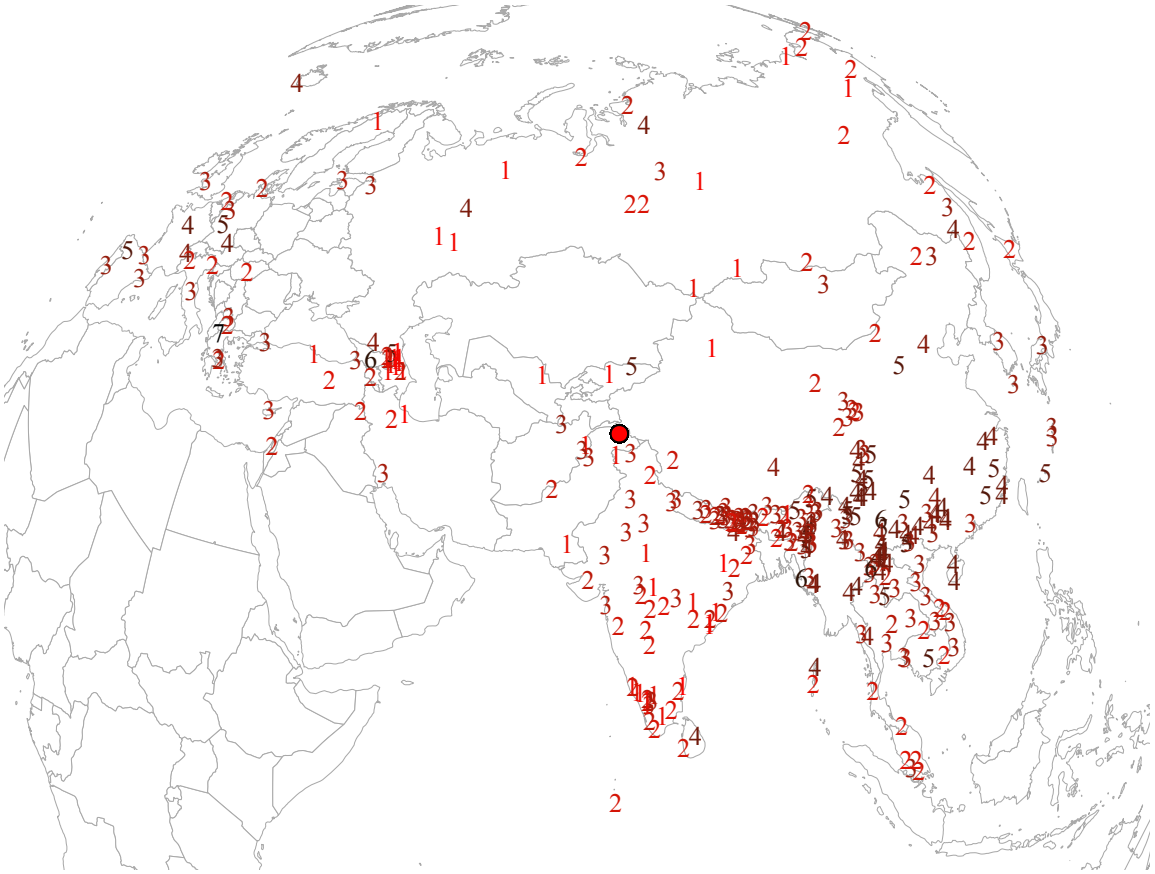


Figure 5.4: Each Eurasian lect's phonological distance from Burushaski, rounded to the closest integer

based on this spatial weight matrix validates the spatial autocorrelation ($p < 0.001$). Finally, using this spatial weight matrix, I perform a spatial regression analysis based on the spatial lag model. The results show that geographical distance to Burushaski is positively correlated with the phonological distance to it ($p < 0.001$). The phonological distance of each lect to Burushaski, rounded to the closest integer, is visualized in Figure 5.4. Based on this observation, we can make the conclusion that phonological distance is positively correlated with geographical distance in Eurasia.

5.2.8 Machine-learning prediction of linguistic areas

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? If machine-learning, when given the phonotactic details of half of the sample lects hypothetically belonging to a phonological area, can well predict whether which sample lects are the other half of the sample lects belonging to the same area, we will be able to conclude that the information provided by Phonotacticon is sufficient to make generalizations on areal similarities, validating the claim that the phonotactic information coded in Phonotacticon can demonstrate phonological area-

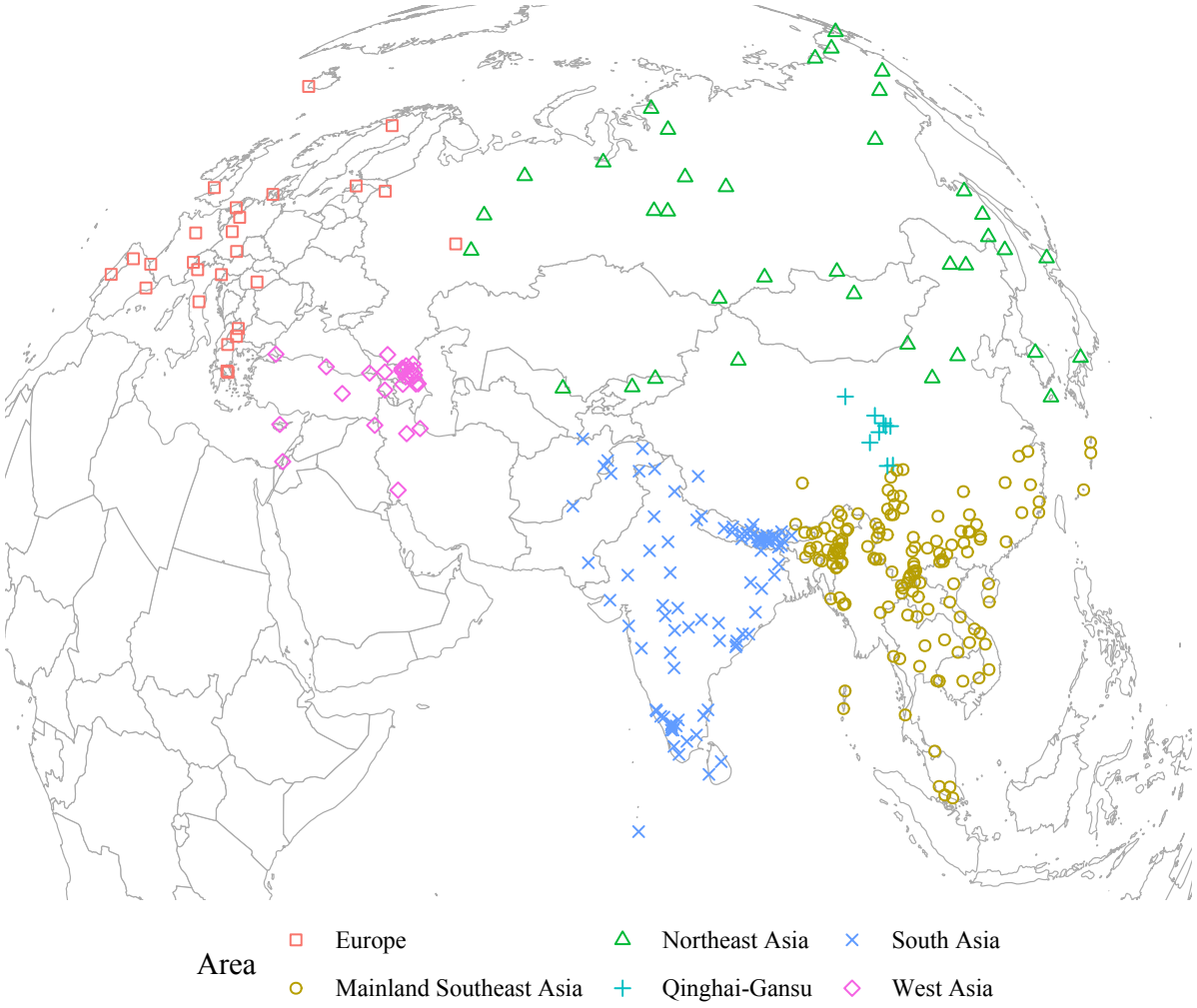


Figure 5.5: Geographically defined regions in Eurasia

hood.

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. While West Asia has not been discussed in this thesis as a potential linguistic area, it is nevertheless posited as a region so that every Eurasian lect will belong to one region. Figure 5.5 visualizes the lects in the predefined seven regions.

The goal is to train a model based on phonological distance to see how well it predicts 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. I also perform the opposite by training the second half and testing it on the first half. Finally, I join the two halves.

Table 5.16 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. The F1 values of individual classes (Table 5.17) show that the model predicts some areas better than others. Mainland Southeast Asia and South Asia are well predicted and Qinghai-Gansu only poorly so. This may be partly because the uneven sample size across different regions, Qinghai-Gansu being the smallest with only nine sample lects. West Asia, although not previously discussed to be a linguistic area, is quite well predicted.

Name	Value
Accuracy	0.61
Kappa	0.47
AccuracyLower	0.55
AccuracyUpper	0.66
AccuracyNull	0.42
AccuracyPValue	0.00
McnemarPValue	0.14

Table 5.16: Confusion matrix based on two halves of Naive Bayes Classifier prediction

Class	F1
Europe	0.43
Mainland Southeast Asia	0.79
Northeast Asia	0.38
Qinghai-Gansu	0.14
South Asia	0.62
West Asia	0.41

Table 5.17: F1 values of individual classes

Figure 5.6 is the visualization of the lects by their predicted areas. Overall, the Naive Bayes Classifier well predicts the six linguistic areas, not including Qinghai-Gansu, which may be due to the small smaple size.

5.3 Comparison with morphosyntactic convergence

I have discussed in Section 2.2 that convergence may be domain-specific: that is, convergence in one domain does not entail convergence in other domains. The phonological convergence patterns we have seen so far do not necessarily imply that similar morphosyntactic convergence has emerged. However, given that phonological convergence is mostly motivated by contact and that lects in contact are likely to converge (although not always), it is feasible that similar morphosyntactic convergence has occurred throughout Eurasia.

In order to test this possibility, I measure the morphosyntactic distance between Eurasian lects using Grambank 1.0 (Skirgård et al. 2023). Grambank 1.0 is a database consisting the

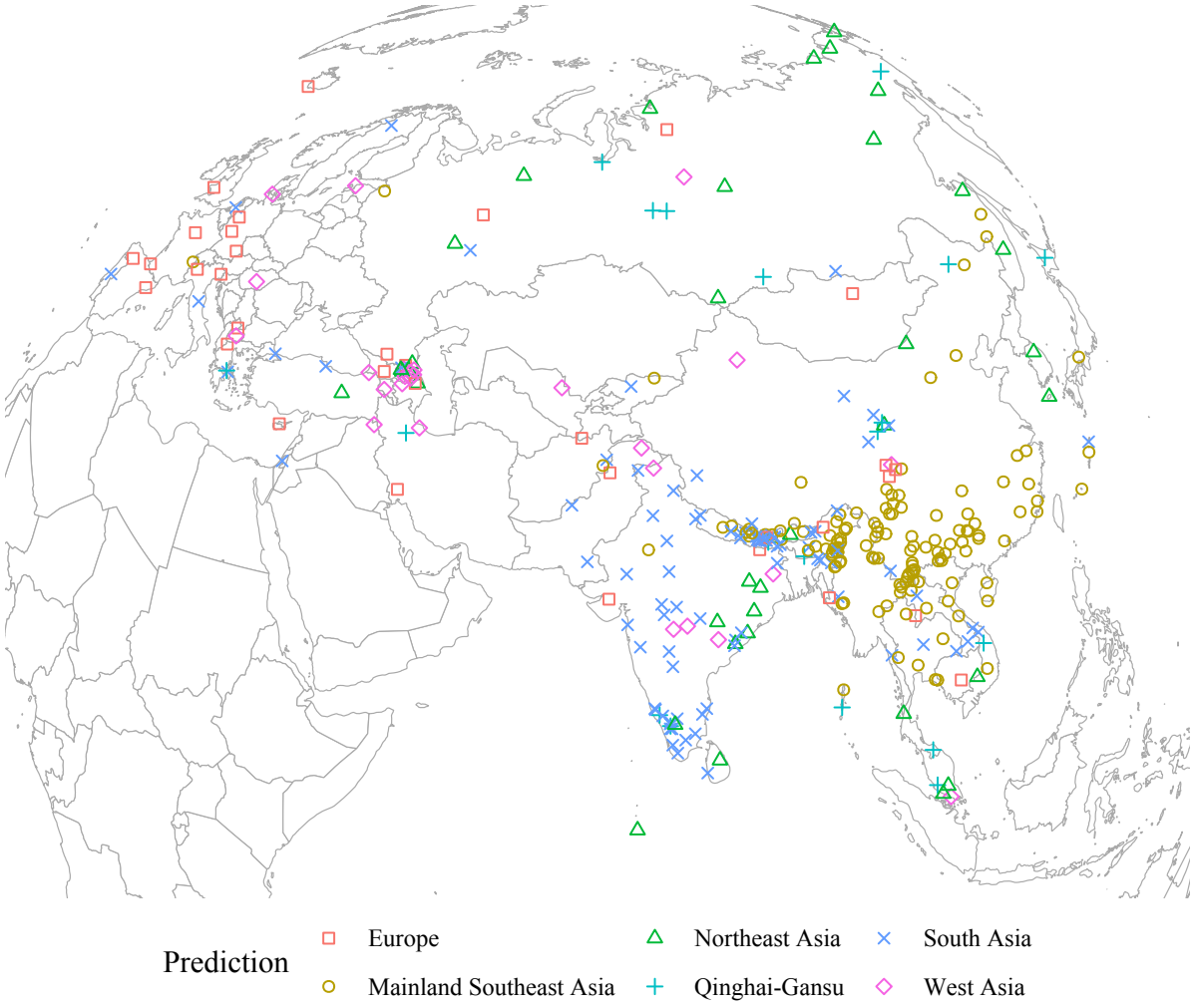


Figure 5.6: Map of predicted linguistic areas

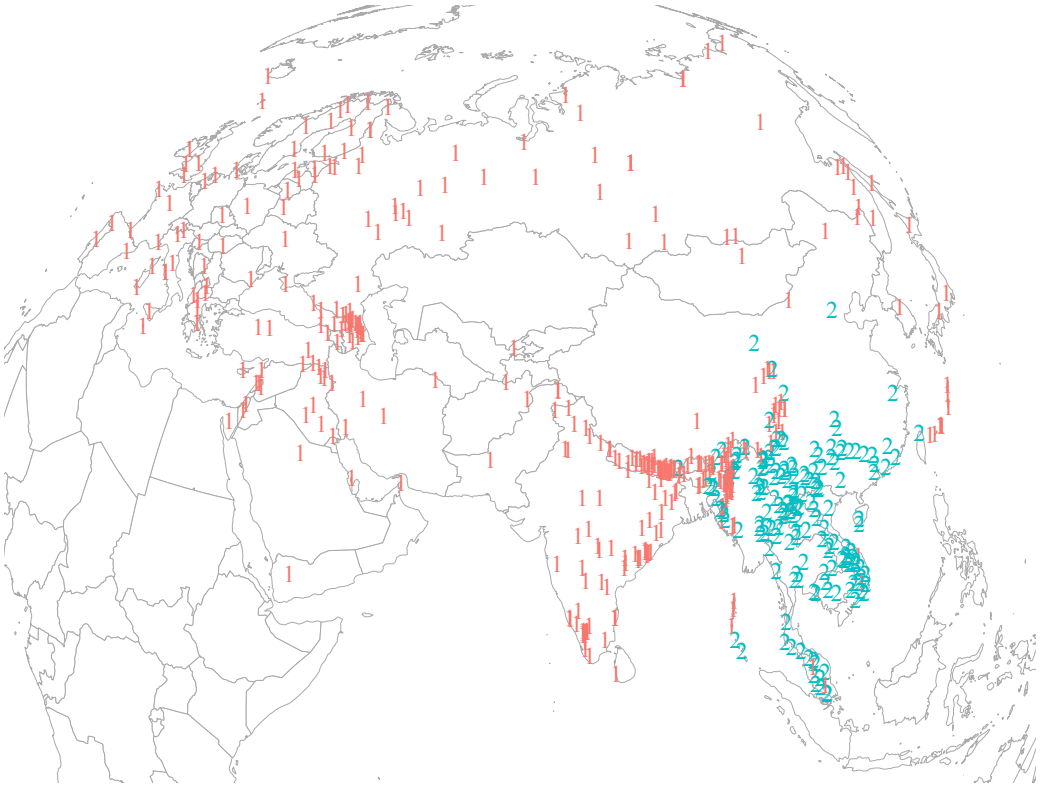


Figure 5.7: Two morphosyntactic clusters of Eurasia

2,457 lects and their values of 195 morphosyntactic features, of which 189 are binary parameters (e.g. *Is there a gender distinction in independent 3rd person pronouns?*). As it does not contain information on phonological features, Phonotacticon can serve as a good complement to Grambank. The two databases used in comparison can cross-validate areal patterns in Eurasia or discover domain-specific patterns.

Based on the 189 binary features of Grambank, I conducted the Manhattan distance between each pair of lects, whose vectors consist of 1 (positive), -1 (negative), or 0 (unknown) values of each feature. (Ca. 13% of all feature values are marked as unknown) Then, based on *k*-means clustering, I clustered the Eurasian lects into two, three, and four clusters, visualized in Figures 5.7–5.9.

It may leave the impression to the reader that the morphosyntactic clusters are somewhat similar to the phonological clusters visualized in Figures 5.1–5.3. In order to test the similarity between the phonological distances and morphosyntactic distances, I ran a Mantel test based on Pearson's correlation coefficient to test the correlation between the two distance matrices. When limiting the sample lects to the intersection between Phonotacticon and Grambank, or 222 lects, the Mantel statistic R is 0.215 from -1 to 1 scale ($p < 0.001$), -1 representing the perfectly negative correlation and 1 the perfectly positive correlation. This shows a a moderate level of correlation between phonological distance and morphosyntactic distance between Eurasian lects. Thus, while phonological similarities and morphosyntactic similarities overlap to some degree, they do not always match and linguistic convergence can be domain-specific.

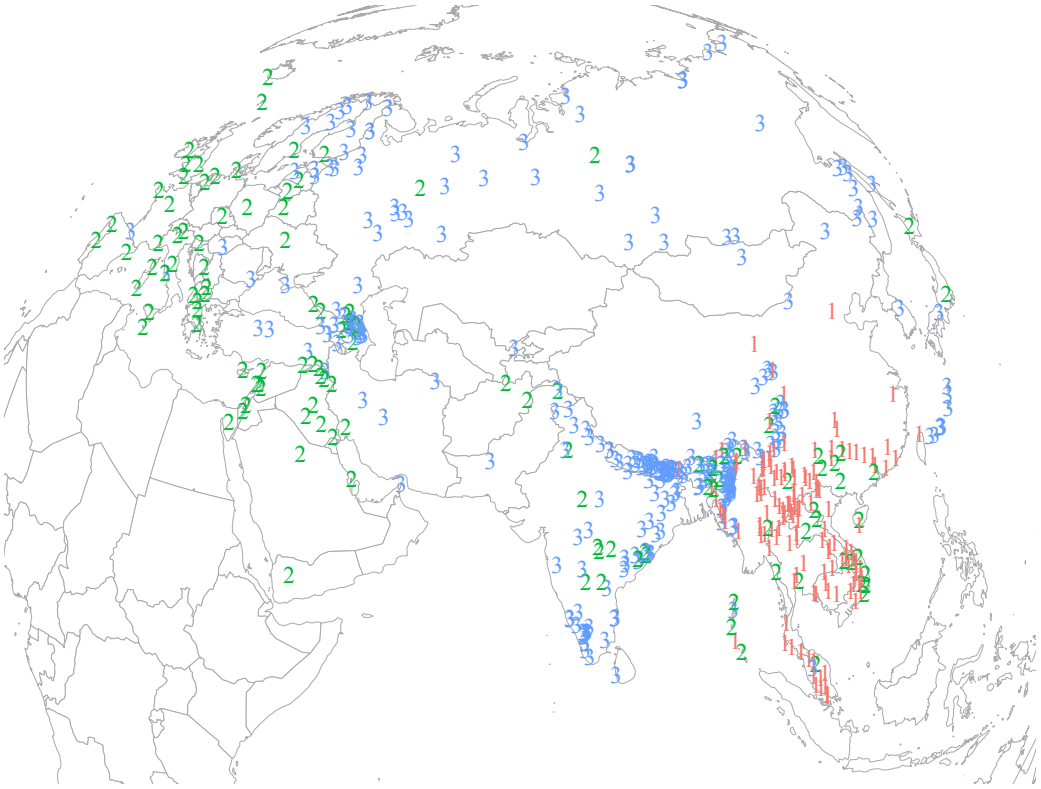


Figure 5.8: Three morphosyntactic clusters of Eurasia

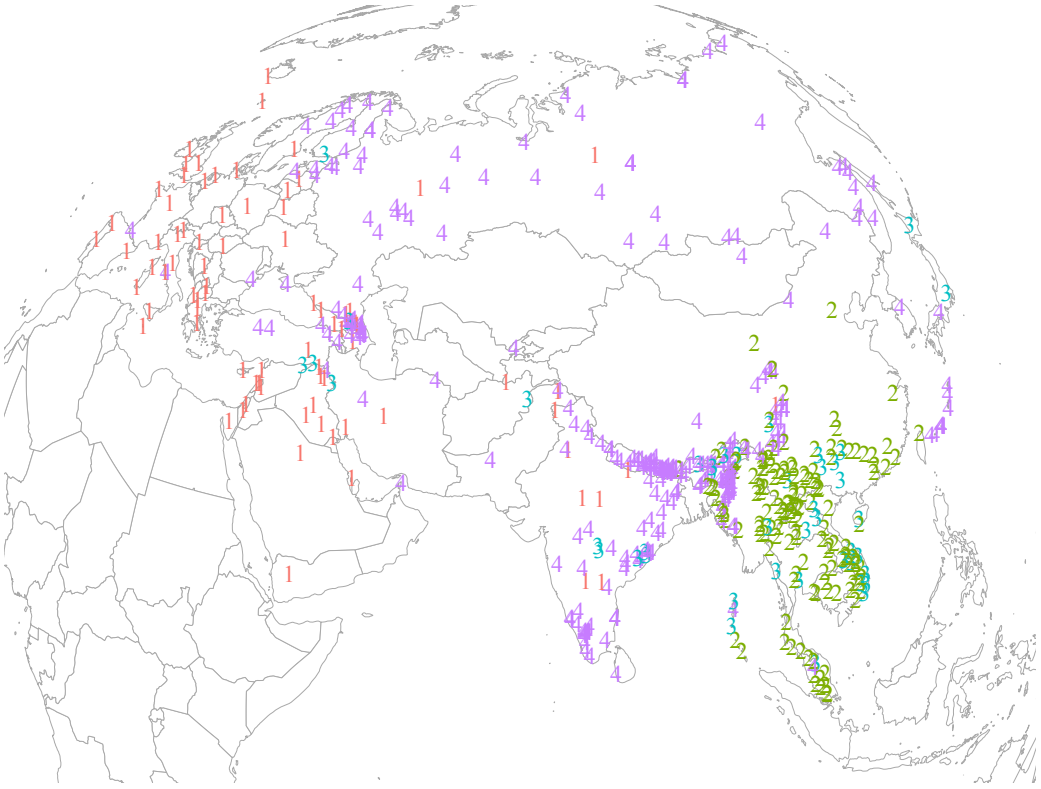


Figure 5.9: Four morphosyntactic clusters of Eurasia

5.4 Comparison with genealogy

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

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

The results shown in Table 5.18 fail the prediction and uphold the null hypothesis. Among the tested Eurasian families, no family shows a statistically significant correlation (FDR = 0.1) between phonological distance and genealogical relatedness. This is true not only for families with a small number of lect pairs, such as Japonic or Hmong-Mien, but also for those with a sizeable sample, such as Indo-European and Sino-Tibetan. This suggests that the phonological distances between Eurasian lects measured via Phonotacticon must have been shaped more strongly by areal contact than by genealogical heritage.

Is the prevalence of areality over genealogy also observable in morphosyntax? In order to make a comparison between the two domains of phonology and morphosyntax, I also tested the correlation between the number of shared layers and morphosyntactic distance measured using Grambank (Section 5.3). Table 5.19 shows the correlation between the number of shared layers and morphosyntactic distances. between the same sample lects. The results are strik-

Family	Number of lect pairs	r	FDR
Nakh-Daghestanian	78	-0.19	1.00
Turkic	66	-0.18	1.00
Tai-Kadai	153	-0.09	1.00
Dravidian	210	-0.07	1.00
Indo-European	1953	-0.03	1.00
Sino-Tibetan	7503	-0.01	1.00
Austroasiatic	351	-0.00	1.00
Mongolic-Khitan	28	0.01	1.00
Uralic	36	0.12	1.00
Japonic	6	0.13	1.00
Tungusic	15	0.34	1.00
Austronesian	21	0.39	1.00
Hmong-Mien	10	0.42	1.00

Table 5.18: Pearson's correlation efficient (r) and the false discovery rate (FDR) between the phonological distance and the number of shared genealogical layers between two lects of the same family

ingly different from those from Phonotacticon (Table 5.18): When corrected for multiple comparison at the False Discovery Rate of 0.1, seven out of fourteen families show a negative correlation between the number of shared layers and morphosyntactic distance. This suggests that morphosyntactic distances may be more heavily related to genealogical heritage whereas phonological distances are more bound to change via contact.

5.5 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 support previously suggested linguistic areas, as we have discussed in Section 2.3. The comparison between the geographical and the phonological distances to Burushaski shows that geographical and phonological distances correlate to some degree in Eurasia. Moreover, The machine learning based on the phonological distances can predict the area of each lect to a moderate degree.

The comparison between the phonological clusters derived from Phonotacticon and the morphosyntactic clusters derived from Grambank (Skirgård et al. 2023) shows that the the convergence patterns in the two domains have only a moderate degree of similarity, suggesting that linguistic convergence may show different areal patterns in different domains. Finally, the absence of correlation between phonological distance and the number of shared layers and

Family	Number of lect pairs	r	FDR
Japonic	15	-0.75	0.01
Uralic	300	-0.57	0.00
Dravidian	435	-0.51	0.00
Indo-European	2016	-0.46	0.00
Turkic	91	-0.37	0.00
Tungusic	36	-0.28	0.57
Sino-Tibetan	17020	-0.27	0.00
Mongolic-Khitan	28	-0.21	0.78
Austroasiatic	3240	-0.19	0.00
Tai-Kadai	120	-0.14	0.63
Nakh-Daghestanian	210	-0.13	0.42
Afro-Asiatic	231	-0.07	0.78
Hmong-Mien	45	0.05	0.78
Austronesian	120	0.12	0.78

Table 5.19: Pearson's correlation efficient (r) and the false discovery rate (FDR) between the morphosyntactic distance and the number of shared genealogical layers between two lects of the same family

the presence of such correlation in morphosyntactic distance suggests that phonology may be more prone to areal convergence than morphosyntax, which may more strongly preserve its genealogical heritage.

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 patterns observable through the phonological distances largely confirm the linguistic areas introduced in Section 2.3.2: Northeast Asia, Qinghai-Gansu, Mainland Southeast Asia, South Asia, and Europe. Thus, we have evidence that these regions form phonological areas, although areahood in other linguistic domains may significantly differ.

One of the limitations of this study based on Phonotacticon is that Phonotacticon does not cover the entirety of a lect's phonotactics, let alone its phonology. It only covers the phonotactic restrictions **within** each of the three slots of a syllable - onset, nucleus, and coda - and not the restrictions **across** these slots, such as which nuclei can follow which onsets. As Phonotacticon is primarily based on segmental information, it also lacks much of suprasegmental information, such as stress patterns. In other words, the information Phonotacticon provides is only a subset of phonological information in its entirety. The phonological convergence patterns and the phonological areas analyzed via Phonotacticon should therefore be interpreted with this limitation in mind.

Moreover, as it is a **phonological** database, it does not cover the **phonetic** characteristics of Eurasian lects, which may also bear areal patterns not reflected in phonological areal patterns. For example, Andrew Hsiu (personal communication) has observed that "pharynx narrowing" may be a phonetic feature of Mainland Southeast Asia, characterizing the voice quality of many speakers of that area. Harnud and Zhou (2021), discussed in Section 2.5.8, also demonstrated the possibility of measuring cross-linguistic distance based on acoustic data, which entails that with sufficient number of sample lects, a quantitative typology of phonetic areal patterns would certainly be possible. Although phonology and phonetics are closely related, inseparable realms (Ohala 1990, cf.), it would certainly be worthy to also investigate phonetic convergence from a typological perspective, which the present study remains agnostic of.

As much as there are limits to Phonotacticon for conducting this study, this study itself is also a limited usage of Phonotacticon, i.e. it does not exploit the full range of possibilities the database entails. In future research, Phonotacticon 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 (2023). "Measuring variation in phoneme inventories". In: *Journal of Language Evolution* (accepted).

Anderson, Gregory D. and David K. Harrison (1999). *Tyvan*. Lincom.

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.

Bakker, Peter (2006). "The Sri Lanka Sprachbund: The Newcomers Portuguese and Malay". In: *Linguistic Areas*. Ed. by Yaron Matras, April McMahon, and Nigel Vincent. Palgrave Macmillan UK, pp. 135–159. DOI: 10.1057/9780230287617_6.

Balanovsky, Oleg, Khadizhat Dibirova, Anna Dybo, Oleg Mudrak, Svetlana Frolova, Elvira Pocheshkhova, Marc Haber, Daniel Platt, Theodore Schurr, Wolfgang Haak, et al. (2011). "Parallel Evolution of Genes and Languages in the Caucasus Region". In: *Molecular Biology and Evolution* 28.10, pp. 2905–2920. DOI: 10.1093/molbev/msr126.

Bauer, Robert S. and Paul K. Benedict (1997). *Modern Cantonese phonology*. Mouton de Gruyter.

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

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Berg, Thomas (1986). "The monophonematic status of diphthongs revisited". In: *Phonetica* 43.4, pp. 198–205.

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.

Boretzky, Norbert (1991). "Contact-induced sound change". In: *Diachronica* 8.1, pp. 1–15. DOI: 10.1075/dia.8.1.02bor.

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.

Catford, J. C. (1977). "Mountain of Tongues: The Languages of the Caucasus". In: *Annual Review of Anthropology* 6.1, pp. 283–314. DOI: 10.1146/annurev.an.06.100177.001435.

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. 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. Cambridge University Press, 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.

— (2008). “Linguistic Diversity in the Caucasus”. In: *Annual Review of Anthropology* 37.1, pp. 131–143. DOI: 10.1146/annurev.anthro.35.081705.123248.

Daniel, Michael and Yury Lander (2011). “The Caucasian languages”. In: *The Languages and Linguistics of Europe*. De Gruyter Mouton, pp. 125–158. DOI: 10.1515/9783110220261.125.

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. (2013). “Order of Subject, Object and Verb”. 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/feature/81A.

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*. Ed. by K. David Harrison, David S. Rood, and Arienne Dwyer. John Benjamins Publishing, 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.

Eliasson, Stig (2022). "The phonological status of Swedish *au* and *eu*: Proposals, evidence, evaluation". In: *Nordic Journal of Linguistics*, pp. 1–42. DOI: 10.1017/s0332586522000233.

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.

Enfield, Nick James and Bernard Comrie (2021). *The Languages of Mainland Southeast Asia*. 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. John Benjamins Publishing Company, pp. 564–591. DOI: 10.1075/cilt.348.26eva.

Fleischer, Jürg and Stephan Schmid (2006). "Zurich German". In: *Journal of the International Phonetic Association* 36.2, pp. 243–253.

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.

Fuchs, Robert (2015). "Word-initial glottal stop insertion, hiatus resolution and linking in British English". In: *Sixteenth annual conference of the international speech communication association*.

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. Brill, pp. 151–168. DOI: 10.1163/9789401206365_011.

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

Goldsmith, John (2011). "The syllable". In: *The handbook of phonological theory*. Ed. by John Goldsmith, Jason Riggle, and Alan C. L Yu. 2nd ed. Wiley, pp. 164–196. DOI: 10.1002/9781444343069.ch6.

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*. Lincom.

Gut, Ulrike (2009). *Introduction to English phonetics and phonology*. Vol. 1. 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*. 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.

Hulst, Harry van der and Nancy A Ritter (1999). "Theories of the syllable". In: *The syllable: views and facts*. Ed. by Harry van der Hulst and Nancy A Ritter. De Gruyter Mouton, pp. 13–52. DOI: 10.1515/9783110806793.13.

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*. Revised. 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.

Janhunen, Juha A (2023). "The unity and diversity of Altaic". In: *Annual Review of Linguistics* 9, pp. 135–154.

Jendraschek, Gerd (2019). "Romance in Contact With Basque". In: *Oxford research encyclopedia of linguistics*. Ed. by Mark Aronoff. DOI: 10.1093/acrefore/9780199384655.013.423.

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

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.

Khabtagaeva, Bayarma (2010). “Mongolic elements in Barguzin Evenki”. In: *Acta Orientalia* 63.1, pp. 9–25.

Klein, Thomas B., E-Ching Ng, and Anthony P. Grant (2020). “Contact-induced change and phonology”. In: *The Oxford handbook of language contact*. Ed. by Anthony P. Grant. Oxford University Press, pp. 74–95. DOI: 10.1093/oxfordhb/9780199945092.013.3.

Kučera, Henry and George K. Monroe (1968). *A comparative quantitative phonology of Russian, Czech, and German*. 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. John Benjamins Publishing Company, pp. 13–38.

Lee 이, Jinho 진호 (2021). *Kwuke umwunlon kanguy* 국어 음운론 강의 *[A course in Korean phonology]*. Jipmundang 집문당.

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

Lǐ, Yuèyuán and Dan Ponsford (2018). “Predicative reduplication: Functions, their relationships and iconicities”. In: *Linguistic Typology* 22.1, pp. 51–117. DOI: 10.1515/lingty-2018-0003.

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

Macklin-Cordes, Jayden L. and Erich R. Round (2020). “Re-evaluating phoneme frequencies”. In: *Frontiers in psychology* 11, p. 570895.

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

— (2013a). “Consonant Inventories”. 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/feature/1A.

— (2013b). “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.

— (2013c). “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*. De Gruyter Mouton. DOI: 10.1515/9783110822816.

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

Matisoff, James A. (2019). "Preface". In: *The Mainland Southeast Asia Linguistic Area*. Ed. by Alice Vittrant and Justin Watkins. De Gruyter Mouton, pp. V–XVI. DOI: 10.1515/9783110401981-202.

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 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.

Nomoto, Hiroki and Hooi Ling Soh (2019). "Malay". In: *The Mainland Southeast Asia Linguistic Area*. Ed. by Alice Vittrant and Justin Watkins. Mouton, pp. 465–522.

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.

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.

Pakendorf, Brigitte (2020). In: *Contact and Siberian languages*. Ed. by Raymond Hickey. Wiley, pp. 669–688. DOI: 10.1002/9781119485094.ch34.

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

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.

Poppe, Nicholas (1972). "On some Mongolian loan words in Evenki". In: *Central Asiatic Journal* 16.2, pp. 95–103.

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 (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. URL: https://www.R-project.org.

Riad, Tomas (2013). *The Phonology of Swedish*. Oxford University Press.

Rubino, Carl (2013). "Reduplication". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Zenodo. DOI: 10.5281/zenodo.7385533.

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.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. (2023). "Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss". In: *Science Advances* 9.16, eadg6175. DOI: 10.1126/sciadv.adg6175.

Slater, Keith W. (2003). *A grammar of Mangghuer: A Mongolic language of China's Qinghai-Gansu sprachbund*. 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.

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

Tamura, Suzuko (2000). *The Ainu language*. 1st ed. Sanseido.

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

Thurgood, Graham, Ela Thurgood, and Li Fengxiang (2014). *A grammatical sketch of Hainan Cham: History, contact, and phonology*. Vol. 643. Walter de Gruyter GmbH & Co KG.

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.

Tuite, Kevin (1999). "The myth of the Caucasian Sprachbund: The case of ergativity". In: *Lingua* 108.1, pp. 1–29. DOI: 10.1016/s0024-3841(98)00037-0.

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 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.

Vogt, Hans (1988). "Substrat et convergence dans l'évolution linguistique. Remarques sur l'évolution et la structure de l'arménien, du géorgien, de l'ossète et du turc". In: *Linguistique caucasienne et arménienne*. Ed. by Hans Vogt. Norwegian University Press, pp. 177–192.

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*. Oxford University Press.

Wu, Hugjiltu (2003). "Bonan". In: *The Mongolic languages*. Ed. by Juha Janhunen. Routledge, pp. 325–345.

Wu, Manxiang (2015). "A grammar of Sanjiang Kam". PhD thesis. University of Hong Kong.

Xie, Yihui, Christophe Dervieux, and Emily Riederer (2020). *R Markdown cookbook*. 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.

Yoshioka, Noboru (2012). "A Reference Grammar of Eastern Burushaski". PhD thesis. Tokyo University of Foreign Studies.

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.

Appendix

The script shown in the following pages is the R script (R Core Team 2024) used to create descriptive visualizations and conduct statistical analyses in this thesis. It is rendered in pdf format via R Markdown (Xie et al. 2020).

(Continued on the next page)

Data

Set up the environment of R Markdown.

```
options(scipen = 100, digits = 3)
options(datatable.print.nrows = 5,
        datatable.print.trunc.cols = T)
```

Load the required R packages.

```
library(data.table)
library(tidytable)
library(stringr)
library(stringi)
library(geosphere)
library(plotly)
library(geodata)
library(plyr)
library(tibble)
library(forcats)
library(purrr)
library(Rfast)
library(e1071)
library(caret)
library(dplyr)
library(anocva)
library(profmem)
library(stargazer)
library(lme4)
library(viridis)
library(spdep)
library(spatialreg)
library(xtable)
library(ggforce)
library(magrittr)
library(vegan)
```

Load Phonotacticon.

```
Phonotacticon <- fread("Phonotacticon1_0.csv") %>%
  as.data.table()

Phonotacticon
```

```
##      Glottocode ISO       Lect   lon  lat          Family         Type
##   1:   aoua1234 aou        A'ou 105.8 26.8       Tai-Kadai         BOOK
##   2:   abaz1241 abq      Abaza  42.0 44.2    Abkhaz-Adyge INCOLLECTION
```

```
##   3:   abkh1244 abk      Abkhaz  41.2 43.1     Abkhaz-Adyge        BOOK
##   4:   adyg1241 ady      Adyghe  39.3 44.0     Abkhaz-Adyge        BOOK
##   5:   akaj1239 akj     Akajeru  93.0 13.2 Great Andamanese        BOOK
##  ---
## 512:   yuec1235 yue Yue Chinese 113.0 23.0     Sino-Tibetan        BOOK
## 513:   zaiw1241 atb       Zaiwa  98.4 24.2     Sino-Tibetan        BOOK
## 514:   zauz1238 zal      Zauzou  98.9 26.5     Sino-Tibetan    PHDTHESIS
## 515:   zbua1234            Zbu 101.7 32.2     Sino-Tibetan    PHDTHESIS
## 516:   zeme1240 nzm    Zeme Naga  93.6 25.3     Sino-Tibetan    PHDTHESIS
## 18 variables not shown: [Author, Year, Title, Publisher, School, Booktitle, Editor, Journal, Pages, Volume, ...]
```

All the subset sample lects are listed below.

```
Eurasia <- Phonotacticon %>%
 .[Onset != '?' &
   Nucleus != '?' &
   Coda != '?' &
   !grepl('C{2,}', Onset) &
   !grepl('C{2,}', Coda) &
   !grepl('\\[.{10}.*?\\]\\[.{10}.*?\\]', Onset) &
   !grepl('\\[.{10}.*?\\]\\[.{10}.*?\\]', Coda)] %>%
 .[, .(Lect, Phoneme, Tone, Onset, Nucleus, Coda)]

Eurasia$Lect
```

```
##   [1] "A'ou"                          "Akajeru"
##   [3] "Amdo Tibetan"                  "Angami Naga"
##   [5] "Ao Naga"                       "Archi"
##   [7] "Aromanian"                     "Arpitan"
##   [9] "Arvanitika Albanian"           "Asho Chin"
##  [11] "Assamese"                      "Asturian-Leonese-Cantabrian"
##  [13] "Atong (India)"                 "Avar"
##  [15] "Baba Malay"                    "Badaga"
##  [17] "Bagvalal"                      "Bantawa"
##  [19] "Basque"                        "Betta Kurumba"
##  [21] "Bezhta"                        "Bih"
##  [23] "Bisu"                          "Biyo"
##  [25] "Bodo-Mech"                     "Bolyu"
##  [27] "Bonan"                         "Budukh"
##  [29] "Bugan"                         "Bujhyal"
##  [31] "Bulo Stieng"                   "Bunan"
##  [33] "Burmese"                       "Burushaski"
##  [35] "Cao Miao"                      "Catalan"
##  [37] "Central Bai"                   "Central Chong"
##  [39] "Central Hongshuihe Zhuang"     "Central Khmer"
##  [41] "Chak"                          "Chintang"
##  [43] "Chitwania Tharu"               "Chong of Chanthaburi"
##  [45] "Chothe"                        "Chukchi"
```

```
##  [47] "Chut"                    "Chuvash"
##  [49] "Cosao"                   "Cypriot Arabic"
##  [51] "Daai Chin"               "Dagur"
##  [53] "Daman-Diu Portuguese"    "Dandami Maria"
##  [55] "Danish"                  "Daohua"
##  [57] "Dari"                    "Darma"
##  [59] "Deori"                   "Dhimal"
##  [61] "Dhivehi"                 "Domari"
##  [63] "Dongxiang"               "Duhumbi"
##  [65] "Dumi"                    "Dungan"
##  [67] "Duoluo Gelao"            "Dutch"
##  [69] "Dzongkha"                "E"
##  [71] "Eastern Katu"            "Eastern Kayah"
##  [73] "Eastern Magar"           "Eastern Newari"
##  [75] "Eastern Panjabi"         "Eastern Tamang"
##  [77] "Enu"                     "Ersu"
##  [79] "Estonian Swedish"        "Evenki"
##  [81] "Forest Enets"            "French"
##  [83] "Friulian"                "Galo"
##  [85] "Gan Chinese"             "Gata'"
##  [87] "Georgian"                "German"
##  [89] "Gheg Albanian"           "Gilaki"
##  [91] "Godoberi"                "Godwari"
##  [93] "Gujarati"                "Gurani"
##  [95] "Hakka Chinese"           "Halbi"
##  [97] "Halh Mongolian"          "Hills Karbi"
##  [99] "Hindi"                   "Hinuq"
## [101] "Hmong Njua"              "Hokkaido Ainu"
## [103] "Honi"                    "Hui Chinese"
## [105] "Hungarian"               "Icelandic"
## [107] "Ingrian"                 "Irula of the Nilgiri"
## [109] "Italian"                 "Iu Mien"
## [111] "Japanese"                "Japhug"
## [113] "Jarawa (India)"          "Jejueo"
## [115] "Jennu Kurumba"           "Jerung"
## [117] "Jinyu Chinese"           "Jiongnai Bunu"
## [119] "Kabardian"               "Kadar"
## [121] "Kado"                    "Kaduo"
## [123] "Kashmiri"                "Kathmandu Valley Newari"
## [125] "Katso"                   "Kayan Lahwi"
## [127] "Kazakh"                  "Kelantan-Pattani Malay"
## [129] "Ket"                     "Khams Tibetan"
## [131] "Khasi"                   "Khezha Naga"
## [133] "Khinalug"                "Khmu"
## [135] "Kirghiz"                 "Kirmanjki"
## [137] "Kman"                    "Kodava"
## [139] "Koi"                     "Koireng"
## [141] "Komi-Zyrian"             "Konda-Dora"
```

```
## [143] "Konkan Marathi"                  "Korean"
## [145] "Korku"                           "Korra Koraga"
## [147] "Koryak"                          "Kotia-Adivasi Oriya-Desiya"
## [149] "Kucong"                          "Kui (India)"
## [151] "Kumaoni"                         "Kumarbhag Paharia"
## [153] "Kumyk"                           "Kurtokha"
## [155] "Kuy"                             "Kyerung"
## [157] "Lachi"                           "Ladino"
## [159] "Lahu"                            "Lak"
## [161] "Lakkia"                          "Lambadi"
## [163] "Lamjung-Melamchi Yolmo"          "Lao"
## [165] "Lashi"                           "Laven"
## [167] "Laz"                             "Leh Ladakhi"
## [169] "Lepcha"                          "Lhomi"
## [171] "Liangmai Naga"                   "Limbu"
## [173] "Lisu"                            "Longchuan Achang"
## [175] "Macedonian"                      "Maithili"
## [177] "Malacca-Batavia Portuguese Creole" "Malavedan"
## [179] "Malayalam"                       "Manchu"
## [181] "Mandarin Chinese"                "Mang"
## [183] "Mangghuer"                       "Manipuri"
## [185] "Mao Naga"                        "Maonan"
## [187] "Maram Naga"                      "Marathi"
## [189] "Marwari (India)"                 "Mewati"
## [191] "Milang"                          "Min Bei Chinese"
## [193] "Min Nan Chinese"                 "Miyako"
## [195] "Mlabri"                          "Modern Greek"
## [197] "Moken"                           "Mon"
## [199] "Mongghul"                        "Moyon"
## [201] "Muduga"                          "Mulam"
## [203] "Mundari"                         "Nanai"
## [205] "Narua"                           "Naukan Yupik"
## [207] "Negidal"                         "Neo-Mandaic"
## [209] "Nepali"                          "Nganasan"
## [211] "Nihali"                          "Nimadi"
## [213] "Nocte Naga"                      "North Azerbaijani"
## [215] "North-Central Dargwa"            "Northeastern Thai"
## [217] "Northern Jinghpaw"               "Northern Pashto"
## [219] "Northern Pinghua"                "Northern Pumi"
## [221] "Northern Thai"                   "Northern Yukaghir"
## [223] "Northwestern Kolami"             "Nung (Myanmar)"
## [225] "Nuristani Kalasha"               "Nyahkur"
## [227] "Odia"                            "Oki-No-Erabu"
## [229] "Oroch"                           "Ostfränkisch"
## [231] "Pa-Hng"                          "Pacoh"
## [233] "Paite Chin"                      "Pela"
## [235] "Peripheral Mongolian"            "Phom Naga"
## [237] "Piemontese"                      "Pite Saami"
```

```
## [239] "Pnar"                    "Pontic"
## [241] "Portuguese"                "Pu-Xian Chinese"
## [243] "Purik-Sham-Nubra"            "Pwo Eastern Karen"
## [245] "Rabha"                   "Rajbanshi"
## [247] "Ravula"                  "Russia Buriat"
## [249] "Russian"                 "Rutul"
## [251] "Sadri"                   "Sadu"
## [253] "Sakha"                   "Sangkong"
## [255] "Sani"                    "Santali"
## [257] "Saurashtra"                "Sedang"
## [259] "Selkup"                  "Semelai"
## [261] "Shixing"                 "Sholaga"
## [263] "Sichuan Yi"                "Sikkimese"
## [265] "Sindhi"                  "Sinhala"
## [267] "Situ"                    "Solu-Khumbu Sherpa"
## [269] "Sora"                    "South Azerbaijani"
## [271] "South Wa"                  "Southeast Pashayi"
## [273] "Southern Altai"              "Southern Amami-Oshima"
## [275] "Southern Jinghpaw"           "Southern Pashto"
## [277] "Southern Pumi"             "Southern Qiang"
## [279] "Southern Rengma Naga"        "Southern Yukaghir"
## [281] "Southwestern Dargwa"         "Sri Lanka Malay"
## [283] "Standard Malay"            "Stau-Dgebshes"
## [285] "Sui"                     "Sunwar"
## [287] "Tai Do-Mene-Yo"            "Tamil"
## [289] "Tangam"                  "Tatar"
## [291] "Thado Chin"                "Thai"
## [293] "Thakali"                 "Thangmi"
## [295] "Thulung"                 "Tibetan"
## [297] "Toda"                    "Tsat"
## [299] "Tsez"                    "Tshangla"
## [301] "Tulu"                    "Tundra Nenets"
## [303] "Tuvinian"                "Udihe"
## [305] "Uighur"                  "Vaagri Booli"
## [307] "Vach-Vasjugan"             "Varhadi-Nagpuri"
## [309] "Vietnamese"                "Waddar"
## [311] "Wambule"                   "Wayu"
## [313] "Welsh"                   "West Yugur"
## [315] "Western Armenian"            "Western Magar"
## [317] "Western Muya"              "Western Ong-Be"
## [319] "Western Parbate Kham"        "Western Puroik"
## [321] "Western Tamang"            "Western Xiangxi Miao"
## [323] "Westphalic"                "Wu Chinese"
## [325] "Wuding-Luquan Yi"            "Wutunhua"
## [327] "Yakkha"                  "Yerong-Southern Buyang"
## [329] "Yongbei Zhuang"            "Youle Jinuo"
## [331] "Yue Chinese"               "Zaiwa"
## [333] "Zauzou"                  "Zbu"
```

```
## [335] "Zeme Naga"
```

Make a list of lects and their geographical coordinates.

```
Lect_LonLat <- Phonotacticon %>%
  .[Lect %in% Eurasia$Lect] %>%
  .[, .(Lect, lon, lat)]

Lect_LonLat
```

```
##             Lect   lon  lat
##   1:        A'ou 105.8 26.8
##   2:     Akajeru  93.0 13.2
##   3: Amdo Tibetan 100.5 34.5
##   4:  Angami Naga  93.9 25.7
##   5:     Ao Naga  94.4 26.3
##  ---
## 331:  Yue Chinese 113.0 23.0
## 332:       Zaiwa  98.4 24.2
## 333:      Zauzou  98.9 26.5
## 334:         Zbu 101.7 32.2
## 335:    Zeme Naga  93.6 25.3
```

Load map data.

```
map <- map_data("world")

head(map)
```

```
##    long  lat group order region subregion
## 1 -69.9 12.5     1     1  Aruba      <NA>
## 2 -69.9 12.4     1     2  Aruba      <NA>
## 3 -69.9 12.4     1     3  Aruba      <NA>
## 4 -70.0 12.5     1     4  Aruba      <NA>
## 5 -70.1 12.5     1     5  Aruba      <NA>
## 6 -70.1 12.6     1     6  Aruba      <NA>
```

Create a map of Eurasia.

```
EurasiaMap <- ggplot(map, aes(x = long, y = lat)) +
  geom_polygon(aes(group = group),
               fill = "white",
               color = "darkgrey",
               size = 0.2) +
  coord_map("ortho",
            orientation = c(20, 70, 0),
```

```
        xlim = c(10, 130),
        ylim = c(0, 90)) +
 theme_void()

EurasiaMap
```

Modified version of PanPhon.

```
PanPhon <- fread("PanPhonPhonotacticon1_0.csv") %>%
 as.data.table() %>%
 unique(by = 'ipa')

PanPhon
```

```
##          ipa syl son cons cont delrel lat nas strid voi sg cg ant cor distr
##    1:     q  -1  -1    1   -1     -1  -1  -1    0  -1 -1 -1  -1  -1     0
##    2:     ɢ  -1  -1    1   -1     -1  -1  -1    0  -1 -1 -1  -1  -1     0
##    3:     ɡ  -1  -1    1   -1     -1  -1  -1    0  -1 -1 -1  -1  -1     0
##    4:     ɢ  -1  -1    1   -1     -1  -1  -1    0   1 -1 -1  -1  -1     0
##    5:     ɢ̣  -1  -1    1   -1     -1  -1  -1    0   1 -1 -1  -1  -1     0
##   ---
## 20241:   ⁿχ̣⁺   1   0    1    1     -1  -1   0    0  -1 -1 -1  -1  -1     0
## 20242:   ⁿɣ̊⁺   1   0    1    1     -1  -1   0    0   1 -1 -1  -1  -1     0
## 20243:    ʔ⁺  -1  -1    1   -1     -1  -1  -1    0  -1 -1 -1  -1  -1     0
```

```
## 20244: ʦʰʷɨ  -1  -1   1  -1    1  -1  -1    0  -1 -1  1   1   1    1
## 20245: ʦʰʷɨ  -1  -1   1  -1    1  -1  -1    0  -1 -1  1   1   1   -1
##        lab
##     1:  -1
##     2:  -1
##     3:  -1
##     4:  -1
##     5:  -1
##    ---
## 20241:  -1
## 20242:  -1
## 20243:  -1
## 20244:  -1
## 20245:  -1
## 7 variables not shown: [hi, lo, back, round, velaric, tense, long]
```

Check if all phonemic transcriptions are present in PanPhon.

```
Transcriptions <- Eurasia$Phoneme %>%
  str_split_fixed(pattern = ' ', n = Inf) %>%
  as.data.table() %>%
  melt(measure.vars = colnames(.)) %>%
  select(-variable) %>%
  filter(value != '') %>%
  distinct() %>%
  mutate(Correct = value %in% PanPhon$ipa)

all(Transcriptions$Correct)
```

```
## [1] TRUE
```

Define a function making a booktabs code.

```
booktabs <- function(x, y) {
  addtorow <- list()
  addtorow$pos <- list(-1, 0, nrow(x))
  addtorow$command <- c('\\toprule ', '\\midrule ', '\\bottomrule ')
  print(x,
        file = y,
        include.rownames = FALSE,
        add.to.row = addtorow,
        hline.after = NULL)
}
```

Sequences of each lect

In this section, I will analyze the sequences of each lect.

Arrange PanPhon segments in alphabetical order.

```
PanPhonOrder <- PanPhon$ipa[
                order(-nchar(PanPhon$ipa),
                   PanPhon$ipa)]

head(PanPhonOrder, 10)
```

```
##  [1] "ʱd̪ɮʲʷ+" "ʱd̪ɮʷː+" "ʱd̪ɮʷʏ+" "ʱd̪ɮʷʔ+" "ʱd̪ʐʲʷ+" "ʱd̪ʐʷː+" "ʱd̪ʐʷʏ+" "ʱd̪ʐʷʔ+"
##  [9] "ʱt̪ɬʲʷ+" "ʱt̪ɬʷː+"
```

Create a regex line of PanPhon in order to split the segments from sequences.

```
PanPhonRegex <- paste0("(?:",
                  paste(PanPhonOrder, collapse="|"),
                 '|B|C|Č|F|G|Ł|L|N|O|P|R|S|T|V|W|X|Z',
                  ")")

str_trunc(PanPhonRegex, 100)
```

```
## [1] "(?:ʱd̪ɮʲʷ|ʱd̪ɮʷː|ʱd̪ɮʷʏ|ʱd̪ɮʷʔ|ʱd̪ʐʲʷ|ʱd̪ʐʷː|ʱd̪ʐʷʏ|ʱd̪ʐʷʔ|ʱt̪ɬ..."
```

Create PanPhon regex including brackets, in order to detect segments within brackets (e. g. [ptk] meaning “/p/, /t/, or /k/”.).

```
PanPhonRegexBrackets <- paste0('(?:',
                          '(?<=\\[).*?(?=\\])|',
                          paste(PanPhonOrder, collapse="|"),
                          '|B|C|Č|F|G|Ł|L|N|O|P|R|S|T|V|W|X|Z',
                          ')')

str_trunc(PanPhonRegexBrackets, 100)
```

```
## [1] "(?:(?<=\\[).*?(?=\\])|ʱd̪ɮʲʷ|ʱd̪ɮʷː|ʱd̪ɮʷʏ|ʱd̪ɮʷʔ|ʱd̪ʐʲʷ|ʱd̪ʐʷː|ʱd̪ʐʷʏ+|..."
```

Define “classes”, i. e. underspecified segments transcribed in capitals (e. g. P for plosives).

```
Classes <- PanPhon %>%
  mutate(B = cons == 1 & lab == 1,
      C = cons == 1,
      Č = cons == 1 & delrel == 1 & son == -1 & cont == -1,
      `F` = cons == 1 & cont == 1 & son == -1,
      G = grepl('j|w|ɥ|ɰ', ipa),
      Ł = cons == 1 & cor == 1 & lat == 1,
      L = cons == 1 & cont == 1 & cor == 1 & son == 1,
      N = nas == 1 & syl == -1,
```

```
      P = cons == 1 & cont == -1 & delrel == -1 & son == -1,
      R = cont == 1 & son == 1 & syl == -1 & !grepl('h|ɦ', ipa),
      S = cons == 1 & cont == 1 & cor == 1 & son == -1,
      `T` = cons == 1 & son != 1,
      V = cons == -1 & cont == 1 & son == 1 & syl == 1,
      W = syl == -1 & voi == 1,
      X = syl == -1 & voi == -1,
      Z = cont == 1 & syl == -1) %>%
 select(ipa, B, C, Č, `F`, G, Ł, L, N, P, R, S, `T`, V, W, X, Z) %>%
 pivot_longer(cols = -ipa,
   names_to = 'Class',
   values_to = 'Value') %>%
 filter(Value) %>%
 select(-Value)

Classes
```

```
## # A tidytable: 82,983 x 2
##    ipa   Class
##    <chr> <chr>
##  1 p     B
##  2 p̪     B
##  3 p̠     B
##  4 b     B
##  5 b̪     B
##  6 b̠     B
##  7 pː    B
##  8 bː    B
##  9 p͡pt   B
## 10 p͡p̪t   B
## # i 82,973 more rows
```

Extract phonemes from the phonemic inventories.

```
Phonemes <- stri_extract_all_regex(Eurasia$Phoneme,
                    pattern = PanPhonRegex,
                    simplify = TRUE) %>%
 as.data.table() %>%
 mutate(Lect = Eurasia$Lect) %>%
 melt(id.vars = 'Lect',
     variable.name = 'Number',
     value.name = 'ipa') %>%
 select(-Number) %>%
 filter(ipa != '')

Phonemes
```

```
##             Lect ipa
```

```
##     1:        A'ou   p
##     2:      Akajeru   p
##     3: Amdo Tibetan   p
##     4:  Angami Naga   p
##     5:      Ao Naga   p
##    ---
## 13089:    Bagvalal  o:
## 13090:    Bagvalal   a
## 13091:    Bagvalal  a:
## 13092:    Bagvalal   ã
## 13093:    Bagvalal  ã:
```

Subset lect, onsets, nuclei, and codas from Phonotacticon.

```
LectONC <- Eurasia %>%
 .[, .(Lect, Onset, Nucleus, Coda)] %>%
 melt(id.vars = 'Lect',
      variable.name = 'Category',
      value.name = 'Sequence')

LectONC
```

```
##              Lect Category
##     1:        A'ou    Onset
##     2:      Akajeru    Onset
##     3: Amdo Tibetan    Onset
##     4:  Angami Naga    Onset
##     5:      Ao Naga    Onset
##    ---
## 1001:  Yue Chinese     Coda
## 1002:       Zaiwa     Coda
## 1003:      Zauzou     Coda
## 1004:         Zbu     Coda
## 1005:    Zeme Naga     Coda
## 1 variable not shown: [Sequence]
```

Extract the sequences from onset, nucleus, and coda categories.

```
Sequences <- LectONC[, tstrsplit(Sequence, ' ', fixed = FALSE)] %>%
 .[, c('Lect', 'Category') := .(LectONC$Lect, LectONC$Category)] %>%
 melt(id.vars = c('Lect', 'Category'),
      variable.name = 'Number',
      value.name = 'Sequence') %>%
 .[, -c('Number')] %>%
 .[!is.na(Sequence)] %>%
 distinct()

Sequences
```

```
##              Lect Category Sequence
##     1:        A'ou    Onset        p
##     2:      Akajeru    Onset        #
##     3: Amdo Tibetan    Onset        #
##     4:  Angami Naga    Onset        #
##     5:      Ao Naga    Onset        #
##    ---
## 28577:      Japhug    Onset       lʁ
## 28578:      Japhug    Onset      ⁀tɕʁ
## 28579:      Japhug    Onset      ⁀tɕʰrʁ
## 28580:      Japhug    Onset       rʁ
## 28581:      Japhug    Onset       jʁ
```

Subset sequences that include underspecified segments (transcribed in capital letters).

```
Capitals <-
 Sequences %>%
 .[grepl('B|C|Č|F|G|Ł|L|N|O|P|R|S|T|V|W|X|Z', Sequence)] %>%
 .[, -c('Category')] %>%
 distinct()

Capitals
```

```
##          Lect Sequence
##   1:    Archi        C
##   2:     Avar        C
##   3: Bagvalal        C
##   4:      Lak        C
##   5:    Rutul        C
##  ---
## 266: Georgian      ČPR
## 267: Georgian      ČFR
## 268: Georgian      FPR
## 269: Georgian      FČR
## 270: Georgian     PFFR
```

Convert the capital letters into the corresponding phonemes in each lect. For example, P ("plosive") in Italian is converted to all the plosive phonemes in Italian phonemic inventory.

```
Decapitalized <-
 stri_extract_all_regex(Capitals$Sequence,
                        pattern = PanPhonRegex,
                        simplify = TRUE) %>%
 as.data.table() %>%
 .[, c('Lect', 'Sequence') :=
     .(Capitals$Lect, Sequence = Capitals$Sequence)] %>%
 melt(id.vars = c('Lect', 'Sequence'),
```

```
      variable.name = 'Order',
     value.name = 'Class') %>%
.[, Order := as.integer(as.factor(Order))] %>%
.[Class != ''] %>%
merge(Classes, all = TRUE, allow.cartesian = TRUE) %>%
.[, ipa := if_else(is.na(ipa), Class, ipa)] %>%
.[, -c('Class')] %>%
merge(Phonemes) %>%
setorder(col = Order) %>%
split(by = c('Lect', 'Sequence')) %>%
lapply(function(x)
       split(x, by = 'Order')) %>%
lapply(function(x)
  lapply(x, function(x)
    x <- x$ipa)) %>%
lapply(function(x)
  expand.grid(x) %>%
    do.call(what = paste0)) %>%
enframe() %>%
unnest() %>%
as.data.table() %>%
separate(col = name,
       into = c('Lect', 'Sequence'),
       sep = '\\.') %>%
setnames('value', 'NewSequence') %>%
merge(Sequences, all = TRUE) %>%
mutate(Sequence =
       if_else(!is.na(NewSequence),
              NewSequence,
              Sequence)) %>%
.[, -c('NewSequence')]
```

Decapitalized

```
## # A tidytable: 45,165 x 3
##    Lect  Sequence Category
##    <chr> <chr>    <fct>
##  1 A'ou  #        Coda
##  2 A'ou  a        Nucleus
##  3 A'ou  ai       Nucleus
##  4 A'ou  aɯ       Nucleus
##  5 A'ou  bl       Onset
##  6 A'ou  d        Onset
##  7 A'ou  e        Nucleus
##  8 A'ou  ei       Nucleus
##  9 A'ou  f        Onset
## 10 A'ou  h        Onset
## # i 45,155 more rows
```

Split the sequences into segments, including bracketed segments (such as [ptk] for "/p/, /t/, or /k/.).

```
ToUnbracket <- stri_extract_all_regex(Decapitalized$Sequence,
                    pattern = PanPhonRegexBrackets,
                    simplify = TRUE) %>%
 as.data.table() %>%
 mutate(Lect = Decapitalized$Lect,
        Category = Decapitalized$Category,
        Sequence = Decapitalized$Sequence) %>%
 melt(id = c('Lect', 'Category', 'Sequence'),
      variable.name = 'Order',
      value.name = 'ipa') %>%
 mutate(Order = Order %>%
          as.factor() %>%
          as.integer()) %>%
 filter(ipa != "")

ToUnbracket
```

```
##              Lect Category Sequence Order ipa
##     1:       A'ou     Coda        #     1   #
##     2:       A'ou  Nucleus        a     1   a
##     3:       A'ou  Nucleus       ai     1   a
##     4:       A'ou  Nucleus       aɯ     1   a
##     5:       A'ou    Onset       bl     1   b
##    ---
## 80877: Ostfränkisch     Coda    ntʃtʃt     6   t
## 80878: Ostfränkisch     Coda  ʁksʰtʃt     6   t
## 80879: Ostfränkisch     Coda  ʁltʃtsʰ     6  sʰ
## 80880: Ostfränkisch     Coda  ʁntʃtsʰ     6  sʰ
## 80881: Ostfränkisch     Coda  ʁpsʰtʃt     6   t
```

Subset bracketed sequences.

```
Bracketed <- ToUnbracket %>%
 filter(grepl('\\[', Sequence))

Bracketed
```

```
##              Lect Category
##  1:        Bezhta     Coda
##  2:  Central Chong    Onset
##  3:   Eastern Katu    Onset
##  4:        Gurani    Onset
##  5:        Kazakh     Coda
## ---
```

```
## 14:        Kazakh     Coda
## 15:         Laven    Onset
## 16:     Piemontese     Coda
## 17:        Thulung    Onset
## 18: Western Puroik    Onset
## 3 variables not shown: [Sequence, Order, ipa]
```

Convert the bracketed sequences into all logically possible sequences. For example, Laven's sequence [bdɟg] [rl] is converted into /br/, /bl/, /dr/, /dl/, /ɟr/ /ɟl/, /gr/, and /gl/.

```
Unbracketed <- Bracketed$ipa %>%
  stri_extract_all_regex(pattern = PanPhonRegex, simplify = TRUE) %>%
  as.data.table() %>%
  mutate(Sequence = Bracketed$Sequence,
         Order = Bracketed$Order) %>%
  melt(id.vars = c('Sequence', 'Order'),
       variable.name = 'Number',
       value.name = 'ipa') %>%
  filter(ipa != '') %>%
  select(-Number) %>%
  setorder(col = Order) %>%
  split(by = 'Sequence') %>%
  lapply(function(x)
    split(x, by = 'Order')) %>%
  lapply(function(x)
    lapply(x, function(x)
      x <- x$ipa)) %>%
  lapply(function(x)
    expand.grid(x) %>%
      do.call(what = paste0)) %>%
  enframe() %>%
  unnest() %>%
  setnames(c('name', 'value'),
           c('Sequence', 'NewSequence')) %>%
  as.data.table()

Unbracketed
```

```
##                                           Sequence NewSequence
##   1: [mnwjlr][pbp’tdt͡’t͡sts’sz͡cç͡cç’ç͡ʝt͡ɬtɬ’ɬkgk’qq’χʁħʕʔhmnwjlr]          mp
##   2: [mnwjlr][pbp’tdt͡’t͡sts’sz͡cç͡cç’ç͡ʝt͡ɬtɬ’ɬkgk’qq’χʁħʕʔhmnwjlr]          np
##   3: [mnwjlr][pbp’tdt͡’t͡sts’sz͡cç͡cç’ç͡ʝt͡ɬtɬ’ɬkgk’qq’χʁħʕʔhmnwjlr]          wp
##   4: [mnwjlr][pbp’tdt͡’t͡sts’sz͡cç͡cç’ç͡ʝt͡ɬtɬ’ɬkgk’qq’χʁħʕʔhmnwjlr]          jp
##   5: [mnwjlr][pbp’tdt͡’t͡sts’sz͡cç͡cç’ç͡ʝt͡ɬtɬ’ɬkgk’qq’χʁħʕʔhmnwjlr]          lp
##  ---
## 468:                                      [pbtdkg][rl]          bl
## 469:                                      [pbtdkg][rl]          tl
## 470:                                      [pbtdkg][rl]          dl
```

```
## 471:                          [pbtdkg][rl]       kl
## 472:                          [pbtdkg][rl]       gl
```

Join the unbracketed sequences into the whole list of sequences. Then split the sequences into segments (e. g. /pl/ into /p/ and /l/).

```
Segments <-
 stri_extract_all_regex(
 Unbracketed$NewSequence,
 pattern = PanPhonRegex,
 simplify = TRUE) %>%
 as.data.table() %>%
 mutate(Sequence = Unbracketed$Sequence,
        NewSequence = Unbracketed$NewSequence) %>%
  pivot_longer(cols = -c(Sequence, NewSequence),
               names_to = 'Order',
               values_to = 'NewIPA') %>%
  mutate(Order = Order %>%
          as.factor() %>%
          as.integer()) %>%
 filter(NewIPA != '') %>%
 full_join(ToUnbracket) %>%
 mutate(Sequence =
          if_else(
            !is.na(NewSequence),
            NewSequence,
            Sequence),
        ipa =
          if_else(
            !is.na(NewIPA),
            NewIPA,
            ipa)) %>%
 select(-NewSequence, -NewIPA) %>%
 as.data.table()
```

```
## Joining with `by = join_by(Sequence, Order)`
```

```
Segments
```

```
##        Sequence Order         Lect Category ipa
##     1:       mp     1      Bezhta     Coda   m
##     2:       np     1      Bezhta     Coda   n
##     3:       wp     1      Bezhta     Coda   w
##     4:       jp     1      Bezhta     Coda   j
##     5:       lp     1      Bezhta     Coda   l
##    ---
## 81803:   ntʃtʃt     6 Ostfränkisch     Coda   t
```

```
## 81804:  ʁksʰtʃt    6 Ostfränkisch    Coda   t
## 81805:  ʁltʃtsʰ    6 Ostfränkisch    Coda  sʰ
## 81806:  ʁntʃtsʰ    6 Ostfränkisch    Coda  sʰ
## 81807:  ʁpsʰtʃt    6 Ostfränkisch    Coda   t
```

Length of sequences

In this section, I will measure the length of each sequence, where length is the number of segments that consist a sequence.

Measure the length of each sequence, in terms of the number of segments involved.

```
Sequences_length <- Segments %>%
  .[, .(Length = max(Order)), by = .(Lect, Category, Sequence)]

Sequences_length
```

```
##           Lect Category Sequence Length
##     1:  Bezhta     Coda       mp      2
##     2:  Bezhta     Coda       np      2
##     3:  Bezhta     Coda       wp      2
##     4:  Bezhta     Coda       jp      2
##     5:  Bezhta     Coda       lp      2
##    ---
## 45619: Zeme Naga    Onset       ŋ      1
## 45620: Zeme Naga  Nucleus       ə      1
## 45621: Zeme Naga  Nucleus      əi      2
## 45622: Zeme Naga  Nucleus      əu      2
## 45623: Zeme Naga    Onset       g      1
```

Join the length of each sequence to segments.

```
Segments <- left_join(Segments, Sequences_length)
```

```
## Joining with `by = join_by(Sequence, Lect, Category)`
```

```
Segments
```

```
##       Sequence Order        Lect Category ipa Length
##    1:       mp     1      Bezhta     Coda   m      2
##    2:       np     1      Bezhta     Coda   n      2
##    3:       wp     1      Bezhta     Coda   w      2
##    4:       jp     1      Bezhta     Coda   j      2
##    5:       lp     1      Bezhta     Coda   l      2
##   ---
```

```
## 81803:   ntʃtʃt     6 Ostfränkisch     Coda   t      6
## 81804:  ʁksʰtʃt     6 Ostfränkisch     Coda   t      6
## 81805:  ʁltʃtsʰ     6 Ostfränkisch     Coda  sʰ      6
## 81806:  ʁntʃtsʰ     6 Ostfränkisch     Coda  sʰ      6
## 81807:  ʁpsʰtʃt     6 Ostfränkisch     Coda   t      6
```

Distance between sequences

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

Count the maximal length of all sequences.

```
MaxLength <- max(Sequences_length$Length)

MaxLength
```

```
## [1] 6
```

Count the number of all the split segments.

```
Segments_number <- nrow(Segments)

Segments_number
```

```
## [1] 81807
```

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 (from 0 to 5).

```
Sequences_rep <- bind_rows(rep(list(Segments), MaxLength))  %>%
  mutate(Position = rep(0:(MaxLength - 1),
                        each = Segments_number)) %>%
  mutate(Order = Order + Position) %>%
  filter(Length + Position <= MaxLength) %>%
  select(-Length)

Sequences_rep
```

```
##        Sequence Order      Lect Category ipa Position
##     1:       mp     1    Bezhta     Coda   m        0
##     2:       np     1    Bezhta     Coda   n        0
##     3:       wp     1    Bezhta     Coda   w        0
##     4:       jp     1    Bezhta     Coda   j        0
```

```
##      5:      lp    1   Bezhta     Coda  l      0
##     ---
## 397684:       z    6 Zeme Naga    Onset  z      5
## 397685:       ŋ    6 Zeme Naga     Coda  ŋ      5
## 397686:       ŋ    6 Zeme Naga    Onset  ŋ      5
## 397687:       ə    6 Zeme Naga  Nucleus  ə      5
## 397688:       g    6 Zeme Naga    Onset  g      5
```

Join segments with their phonological features (retrieved from PanPhon). Each feature is assigned the value of the position.

```
Sequences_features <- Sequences_rep %>%
 left_join(PanPhon, by = 'ipa') %>%
           melt(id = c('Lect',
                  'Category',
                  'Sequence',
                  'Order',
                  'ipa',
                  'Position'),
           variable.name = 'Feature',
           value.name = 'Value') %>%
 mutate(Feature = paste0(Feature, Order)) %>%
 dcast(Lect + Category + Sequence + Position ~ Feature,
      value.var = 'Value',
      fun.aggregate = sum,
      fill = 0) %>%
 mutate(SequencePosition = paste0(Sequence, Position)) %>%
 select(-Lect, -Category, -Position, -Sequence) %>%
 distinct()

Sequences_features
```

```
##       ant1 ant2 ant3 ant4 ant5 ant6 back1 back2 back3 back4 back5 back6 cg1
##    1:    1    1    0    0    0    0    -1    -1    0    0    0    0  -1
##    2:    0    1    1    0    0    0    0    -1    -1    0    0    0   0
##    3:    0    0    1    1    0    0    0    0    -1    -1    0    0   0
##    4:    0    0    0    1    1    0    0    0    0    -1    -1    0   0
##    5:    0    0    0    0    1    1    0    0    0    0    -1    -1   0
##   ---
## 60400:    0   -1    0    0    0    0    0    1    0    0    0    0   0
## 60401:    0    0   -1    0    0    0    0    0    1    0    0    0   0
## 60402:    0    0    0   -1    0    0    0    0    0    1    0    0   0
## 60403:    0    0    0    0   -1    0    0    0    0    0    1    0   0
## 60404:    0    0    0    0    0   -1    0    0    0    0    0    1   0
##       cg2
##    1:  -1
##    2:  -1
##    3:   0
```

```
##     4:   0
##     5:   0
##    ---
## 60400:  -1
## 60401:   0
## 60402:   0
## 60403:   0
## 60404:   0
## 119 variables not shown: [cg3, cg4, cg5, cg6, cons1, cons2, cons3, cons4, cons5, cons6, ...]
```

Calculate the Saporta distance between each pair of sequences.

```
Sequences_distance <- Sequences_features %>%
  select(-SequencePosition) %>%
  Dist(method = 'manhattan') %>%
  as.data.table()

Sequences_distance[1:10, 1:10]
```

```
##     V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
##  1:  0 50 78 78 78 25 25 59 59  59
##  2: 50  0 50 78 78 59 25 25 59  59
##  3: 78 50  0 50 78 59 59 25 25  59
##  4: 78 78 50  0 50 59 59 59 25  25
##  5: 78 78 78 50  0 59 59 59 59  25
##  6: 25 59 59 59 59  0 40 40 40  40
##  7: 25 25 59 59 59 40  0 40 40  40
##  8: 59 25 25 59 59 40 40  0 40  40
##  9: 59 59 25 25 59 40 40 40  0  40
## 10: 59 59 59 25 25 40 40 40 40   0
```

In order to name the rows and the columns of the distance matrix, create a vector of all sequences in different positions.

```
SequenceVectors <-
  str_replace(Sequences_features$SequencePosition, '[0-9]', '')

head(SequenceVectors, 10)
```

```
##  [1] "bl" "bl" "bl" "bl" "bl" "d"  "d"  "d"  "d"  "d"
```

Use this vector to name the rows and the columns of the distance matrix.

```
Sequences_distance[, Sequence := SequenceVectors]

setcolorder(Sequences_distance,
```

```
        c(ncol(Sequences_distance), 1:(ncol(Sequences_distance) - 1)))

setnames(Sequences_distance, c('Sequence', Sequences_features$SequencePosition))

Sequences_distance[1:10, 1:10]
```

```
##    Sequence bl0 bl1 bl2 bl3 bl4 d0 d1 d2 d3
##  1:       bl   0  50  78  78  78 25 25 59 59
##  2:       bl  50   0  50  78  78 59 25 25 59
##  3:       bl  78  50   0  50  78 59 59 25 25
##  4:       bl  78  78  50   0  50 59 59 59 25
##  5:       bl  78  78  78  50   0 59 59 59 59
##  6:        d  25  59  59  59  59  0 40 40 40
##  7:        d  25  25  59  59  59 40  0 40 40
##  8:        d  59  25  25  59  59 40 40  0 40
##  9:        d  59  59  25  25  59 40 40 40  0
## 10:        d  59  59  59  25  25 40 40 40 40
```

As I have shown above, I need to choose the minimal distance between two sequences mapped onto each other in different positions. Thus I calculate the minimal distance per each sequence pair.

Make a vector of all sequences.

```
AllSequences <- unique(Segments$Sequence)

head(AllSequences)
```

```
## [1] "mp" "np" "wp" "jp" "lp" "rp"
```

Calculate the minimal distance of every sequence to each sequence.

```
for (i in AllSequences){

 csv_name <- paste0('~/Phonotacticon/Sequences/', i, '.csv')

 OneSequence <-
   Sequences_distance[Sequence == i] %>%
   melt(id.vars = 'Sequence',
        variable.name = 'i.Sequence',
        value.name = 'Distance') %>%
   .[, i.Sequence := gsub('[0-9]', '', i.Sequence)] %>%
   .[, .(Distance = min(Distance)),
    by = .(i.Sequence)] %>%
   .[, Sequence := i]

 fwrite(OneSequence, csv_name)
}
```

Make a list of the sequence file names.

```
Sequences_file_list <-
  list.files(path = '~/Phonotacticon/Sequences',
             pattern = '.*')

head(Sequences_file_list)
```

```
## [1] "#.csv" "a.csv" "ă.csv" "ã.csv" "a̰.csv" "ą̰.csv"
```

Read them into a data table.

```
Sequences_MinDistance <-
  map_df(Sequences_file_list, ~ fread(paste0('~/Phonotacticon/Sequences/', .x))) %>%
  as.data.table()

Sequences_MinDistance
```

```
##            i.Sequence Distance Sequence
##         1:         bl       39        #
##         2:          d       20        #
##         3:          f       20        #
##         4:          h       19        #
##         5:          j       19        #
##        ---
## 187361340:       ⁿdzw       27       XX
## 187361341:       ⁿdʑw       27       XX
## 187361342:        ⁿŋg       16       XX
## 187361343:        ⁿɻw       22       XX
## 187361344:        ⁿgʷ       29       XX
```

Sequences that are the most similar to /pl/.

```
pl <- Sequences_MinDistance %>%
  filter(Sequence == 'pl') %>%
  arrange(Distance)

pl %>%
  .[1:20] %>%
  .[, Sequence := NULL] %>%
  setnames(old = 'i.Sequence', new = 'Sequence') %>%
  xtable(type = 'latex',
         label = 'pl',
         caption = 'Sequences the most similar to /pl/') %>%
  booktabs('pl.tex')

pl
```

```
##       i.Sequence Distance Sequence
##    1:         pl        0       pl
##    2:        pˣl        1       pl
##    3:         bl        2       pl
##    4:        pʰl        2       pl
##    5:         p̥l        2       pl
##   ---
## 13684:    ltsʰtʃt       93       pl
## 13685:    ntsʰtʃt       93       pl
## 13686:    ʁksʰtʃt       93       pl
## 13687:     ltʃtʃt       95       pl
## 13688:     ntʃtʃt       95       pl
```

Sequences that are the most similar to /ia/.

```
ia <- Sequences_MinDistance %>%
  filter(Sequence == 'ia') %>%
  arrange(Distance)

ia %>%
  .[1:20] %>%
  .[, Sequence := NULL] %>%
  setnames(old = 'i.Sequence', new = 'Sequence') %>%
  xtable(type = 'latex',
         label = 'ia',
         caption = 'Sequences the most similar to /ia/') %>%
  booktabs('ia.tex')

ia
```

```
##       i.Sequence Distance Sequence
##    1:         ia        0       ia
##    2:         iă        0       ia
##    3:         iæ        0       ia
##    4:         ie        2       ia
##    5:         ea        2       ia
##   ---
## 13684:    ltsʰtʃt      110       ia
## 13685:     ltʃtʃt      110       ia
## 13686:    ʁksʰtʃt      110       ia
## 13687:    ntsʰtʃt      112       ia
## 13688:     ntʃtʃt      112       ia
```

Distance between lects

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

Redefine sequences with new (decapitalized and unbracketed) sequences.

```
NewSequences <- Segments %>%
  select(Lect, Category, Sequence)

NewSequences
```

```
##              Lect Category Sequence
##     1:      Bezhta     Coda       mp
##     2:      Bezhta     Coda       np
##     3:      Bezhta     Coda       wp
##     4:      Bezhta     Coda       jp
##     5:      Bezhta     Coda       lp
##    ---
## 81803: Ostfränkisch     Coda    ntʃtʃt
## 81804: Ostfränkisch     Coda  ʁksʰtʃt
## 81805: Ostfränkisch     Coda  ʁltʃtsʰ
## 81806: Ostfränkisch     Coda  ʁntʃtsʰ
## 81807: Ostfränkisch     Coda  ʁpsʰtʃt
```

Split the data.table of sequences into separate lects.

```
for (i in Eurasia$Lect){
  csv_name <- paste0('~/Phonotacticon/Lects/', i, '.csv')

  Lect <- NewSequences[Lect == i]

  fwrite(Lect, csv_name)
}
```

Assign every onset/nucleus/coda sequence to every other onset/nucleus/coda sequence.

```
for (i in Eurasia$Lect){
  csv_name <- paste0('~/Phonotacticon/Lects/', i, '.csv')

  Lect <- fread(csv_name) %>%
    as.data.table() %>%
    .[Sequences_MinDistance, on = .(Sequence), nomatch = 0]

  fwrite(Lect, csv_name)
}
```

Calculate the minimal distance between the onset/nucleus/coda inventory of each pair of lects.

```
for (i in Eurasia$Lect){

  csv_name <- paste0('~/Phonotacticon/Lects/', i, '.csv')
```

```
  Lect <-
    fread(csv_name) %>%
    as.data.table() %>%
    .[NewSequences,
      on = .(Category,
             i.Sequence = Sequence),
      allow.cartesian = TRUE] %>%
    .[, .(Distance = min(Distance)),
      by = .(i.Lect, Sequence, Category)] %>%
    .[, .(Distance = mean(Distance)),
        by = .(i.Lect, Category)]

  Lect$Lect <- i

  fwrite(Lect, csv_name)

}
```

Make a list of .csv file names.

```
Lects_file_list <-
  list.files(path = '~/Phonotacticon/Lects',
             pattern = '.*')

head(Lects_file_list)
```

```
## [1] "A'ou.csv"        "Akajeru.csv"     "Amdo Tibetan.csv" "Angami Naga.csv"
## [5] "Ao Naga.csv"     "Archi.csv"
```

Read the .csv files into a data.table.

```
ONC_distance_mean <-
  map_df(Lects_file_list, ~ fread(paste0('~/Phonotacticon/Lects/', .x))) %>%
  as.data.table()

ONC_distance_mean
```

```
##              i.Lect Category Distance      Lect
##      1:      Bezhta     Coda     1.33      A'ou
##      2: Central Chong    Onset     2.11      A'ou
##      3:  Eastern Katu    Onset     2.39      A'ou
##      4:       Gurani    Onset     2.81      A'ou
##      5:       Kazakh     Coda     0.00      A'ou
##     ---
## 336671:          Zbu     Coda     1.29 Zeme Naga
## 336672:          Zbu  Nucleus     9.67 Zeme Naga
```

```
## 336673:    Zeme Naga    Onset    0.00 Zeme Naga
## 336674:    Zeme Naga     Coda    0.00 Zeme Naga
## 336675:    Zeme Naga  Nucleus    0.00 Zeme Naga
```

In order to create a data.table of lect pairs, make a dummy column.

```
LectDummy <- Lect_LonLat %>%
  .[, Dummy := 'Dummy']

LectDummy
```

Create a data.table of lect pairs.

```
Lect_vs_Lect <- LectDummy %>%
  .[LectDummy, on = 'Dummy', allow.cartesian = TRUE] %>%
  .[, Lect_vs_Lect := str_c(pmin(as.character(Lect), as.character(i.Lect)),
                            'vs.',
                            pmax(as.character(Lect), as.character(i.Lect)),
                           sep = ' ')] %>%
  .[, -c('Dummy')]

Lect_vs_Lect
```

```
##              Lect   lon  lat    i.Lect i.lon i.lat            Lect_vs_Lect
##      1:      A'ou 105.8 26.8      A'ou 105.8  26.8           A'ou vs. A'ou
##      2:   Akajeru  93.0 13.2      A'ou 105.8  26.8        A'ou vs. Akajeru
##      3: Amdo Tibetan 100.5 34.5    A'ou 105.8  26.8    A'ou vs. Amdo Tibetan
##      4:  Angami Naga  93.9 25.7    A'ou 105.8  26.8     A'ou vs. Angami Naga
##      5:     Ao Naga  94.4 26.3     A'ou 105.8  26.8        A'ou vs. Ao Naga
##     ---
## 112221:  Yue Chinese 113.0 23.0 Zeme Naga  93.6  25.3 Yue Chinese vs. Zeme Naga
## 112222:       Zaiwa  98.4 24.2 Zeme Naga  93.6  25.3       Zaiwa vs. Zeme Naga
## 112223:      Zauzou  98.9 26.5 Zeme Naga  93.6  25.3      Zauzou vs. Zeme Naga
## 112224:         Zbu 101.7 32.2 Zeme Naga  93.6  25.3         Zbu vs. Zeme Naga
## 112225:   Zeme Naga  93.6 25.3 Zeme Naga  93.6  25.3   Zeme Naga vs. Zeme Naga
```

Assign the lect pair column to the mean distances and then detect the maximal distance between the Lect A vs. Lect B pair and the Lect B vs. Lect A pair. The result is the onset/nucleus/coda distance between each pair of lects.

```
ONC_distance <-
  ONC_distance_mean %>%
  .[Lect_vs_Lect, on = .(Lect, i.Lect)] %>%
  .[, .(Lect_vs_Lect, Category, Distance)] %>%
  .[, .(Distance = max(Distance)),
      by = .(Lect_vs_Lect, Category)] %>%
    dcast(., Lect_vs_Lect ~ Category, value.var = 'Distance') %>%
```

```
 relocate(Lect_vs_Lect, Onset, Nucleus, Coda)

ONC_distance %>%
 sample_n(5) %>%
 xtable(type = 'latex',
        label = 'ONCsample',
        caption = 'Five random lect pairs and their onset, nucleus, and coda distances') %>%
 booktabs('ONC.tex')

ONC_distance
```

```
##                Lect_vs_Lect Onset Nucleus  Coda
##     1:          A'ou vs. A'ou  0.00    0.00  0.00
##     2:       A'ou vs. Akajeru  3.83    4.72  5.59
##     3:  A'ou vs. Amdo Tibetan  8.40   20.80  5.00
##     4:   A'ou vs. Angami Naga  2.42   10.36 13.00
##     5:       A'ou vs. Ao Naga  3.83   10.20  5.00
##    ---
## 56276:         Zauzou vs. Zbu 19.48    9.63 17.73
## 56277:   Zauzou vs. Zeme Naga  2.08    4.41 16.57
## 56278:           Zbu vs. Zbu  0.00    0.00  0.00
## 56279:      Zbu vs. Zeme Naga 19.89    9.67  3.00
## 56280: Zeme Naga vs. Zeme Naga  0.00    0.00  0.00
```

Distance of tones

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

Count the number of tones in each lect.

```
Tones <- Eurasia %>%
 .[, .(Lect, Tone)] %>%
 .[, Tone := gsub("\\-", NA, Tone)] %>%
 .[, Tone := str_count(Tone, " ") + 1]

Tones[is.na(Tones$Tone),]$Tone <- 0

Tones
```

```
##            Lect Tone
##   1:        A'ou    4
##   2:     Akajeru    0
##   3: Amdo Tibetan    0
##   4:  Angami Naga    5
##   5:      Ao Naga    3
##  ---
```

```
## 331:  Yue Chinese    6
## 332:       Zaiwa    4
## 333:      Zauzou    6
## 334:         Zbu    2
## 335:   Zeme Naga    2
```

Calculate the Canberra distance between the numbers of tones of each pair of lects.

```
Tones_distance <- Tones %>%
 .[, -c('Lect')] %>%
 dist(method = 'canberra') %>%
 as.matrix() %>%
 as.data.table() %>%
 setnames(Tones$Lect) %>%
 .[, Lect := Tones$Lect] %>%
 melt(id = 'Lect',
      variable.name = 'Lect2',
      value.name = 'Tone') %>%
 .[, Tone := replace_na(Tone, 0)] %>%
 .[, Lect_vs_Lect := str_c(pmin(as.character(Lect), as.character(Lect2)),
                           'vs.',
                           pmax(as.character(Lect), as.character(Lect2)),
                          sep = ' ')] %>%
 .[, .(Lect_vs_Lect, Tone)] %>%
 distinct()

Tones_distance
```

```
##                Lect_vs_Lect  Tone
##     1:          A'ou vs. A'ou 0.000
##     2:       A'ou vs. Akajeru 1.000
##     3:  A'ou vs. Amdo Tibetan 1.000
##     4:   A'ou vs. Angami Naga 0.111
##     5:       A'ou vs. Ao Naga 0.143
##    ---
## 56276:         Zauzou vs. Zbu 0.500
## 56277:   Zauzou vs. Zeme Naga 0.500
## 56278:           Zbu vs. Zbu 0.000
## 56279:      Zbu vs. Zeme Naga 0.000
## 56280: Zeme Naga vs. Zeme Naga 0.000
```

Join segmental distance with tonal distance and normalize the four distances.

```
ONCT_distance <- ONC_distance %>%
 full_join(Tones_distance) %>%
 select(-Lect_vs_Lect) %>%
 scale() %>%
```

```
  as.data.table() %>%
  mutate(Lect_vs_Lect = ONC_distance$Lect_vs_Lect) %>%
  relocate(Lect_vs_Lect)
```

```
## Joining with `by = join_by(Lect_vs_Lect)`
```

```
ONCT_distance
```

```
##                Lect_vs_Lect   Onset Nucleus   Coda    Tone
##     1:          A'ou vs. A'ou -1.5333  -1.487 -1.266 -1.1111
##     2:       A'ou vs. Akajeru -0.8563  -0.450 -0.597  1.0056
##     3:  A'ou vs. Amdo Tibetan -0.0497   3.079 -0.667  1.0056
##     4:   A'ou vs. Angami Naga -1.1055   0.787  0.291 -0.8759
##     5:       A'ou vs. Ao Naga -0.8563   0.752 -0.667 -0.8088
##    ---
## 56276:         Zauzou vs. Zbu  1.9078   0.627  0.858 -0.0528
## 56277:   Zauzou vs. Zeme Naga -1.1665  -0.520  0.719 -0.0528
## 56278:           Zbu vs. Zbu -1.5333  -1.487 -1.266 -1.1111
## 56279:      Zbu vs. Zeme Naga  1.9803   0.635 -0.907 -1.1111
## 56280: Zeme Naga vs. Zeme Naga -1.5333  -1.487 -1.266 -1.1111
```

Overall distance

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).

```
PhonoDist <- ONCT_distance %>%
  mutate(Distance = sqrt((Onset - min(Onset)) ^ 2 +
                         (Nucleus - min(Nucleus)) ^ 2 +
                         (Coda - min(Coda)) ^ 2 +
                         (Tone - min(Tone)) ^ 2)) %>%
  select(Lect_vs_Lect, Distance)

PhonoDist %>%
  arrange(Distance) %>%
  filter(Distance > 0) %>%
  head(10) %>%
  xtable(caption = 'The ten lect pairs with the shortest phonological distance',
         label = 'FirstTenOverall',
         type = 'latex') %>%
  booktabs('PhonoDist.tex')

fwrite(PhonoDist, 'PhonoDist.csv')

PhonoDist
```

```
##                Lect_vs_Lect Distance
##     1:          A'ou vs. A'ou     0.00
##     2:       A'ou vs. Akajeru     2.54
##     3:   A'ou vs. Amdo Tibetan     5.28
##     4:    A'ou vs. Angami Naga     2.80
##     5:       A'ou vs. Ao Naga     2.43
##    ---
## 56276:         Zauzou vs. Zbu     4.68
## 56277:    Zauzou vs. Zeme Naga     2.48
## 56278:           Zbu vs. Zbu     0.00
## 56279:      Zbu vs. Zeme Naga     4.12
## 56280: Zeme Naga vs. Zeme Naga     0.00
```

Ten selected lects.

```
TenLects <- c('Basque',
              'Evenki',
              'Georgian',
              'Hindi',
              'Japanese',
              'Kazakh',
              'Mandarin Chinese',
              'Sri Lanka Malay',
              'Standard Malay',
              'Tsat')

TenLects
```

```
##  [1] "Basque"          "Evenki"          "Georgian"        "Hindi"
##  [5] "Japanese"        "Kazakh"          "Mandarin Chinese" "Sri Lanka Malay"
##  [9] "Standard Malay"   "Tsat"
```

The distances of other lects to each of the ten selected lects.

```
TenLectsDist <- PhonoDist %>%
 .[Lect_vs_Lect, on = .(Lect_vs_Lect)] %>%
 filter(Lect %in% TenLects,
        Lect != i.Lect) %>%
 .[Phonotacticon, on = .(i.Lect = Lect), nomatch = 0] %>%
 select(Family, Lect, i.Lect, Distance) %>%
 arrange(Distance)

TenLectsDist
```

```
##                Family          Lect               i.Lect Distance
##    1:      Sino-Tibetan Standard Malay         Atong (India)    0.262
##    2:      Austronesian Standard Malay Kelantan-Pattani Malay    0.362
```

```
##    3:            Uralic Standard Malay         Tundra Nenets    0.389
##    4:       Jarawa-Onge Standard Malay        Jarawa (India)    0.403
##    5: Chukotko-Kamchatkan Standard Malay               Koryak    0.411
##   ---
## 3336:       Sino-Tibetan       Georgian         Jinyu Chinese    8.715
## 3337:       Sino-Tibetan       Georgian          Paite Chin    8.720
## 3338:    Mongolic-Khitan       Georgian        Halh Mongolian    8.733
## 3339:       Sino-Tibetan       Georgian        Hakka Chinese    8.823
## 3340:       Sino-Tibetan       Georgian                 Lahu    8.950
```

Write each of the ten lects and their twenty closest lects as .tex files.

```
for (i in TenLects){
  TwentyLects <-
    TenLectsDist[Lect == i] %>%
    .[, Lect := NULL] %>%
    setnames(., 'i.Lect', 'Lect') %>%
    head(20)

  Caption <- paste0('Twenty lects closest to ', i)
  Tex <- paste0(Caption, '.tex')

  TwentyLects %>%
    xtable(label = i, caption = Caption, style = 'latex') %>%
    booktabs(Tex)
}
```

Clustering the lects

Based on these distances, I cluster similar lects together and detect areal patterns.

Create a data.table consisting only of phonological distances.

```
PhonoDistNumbers <- PhonoDist %>%
  .[Lect_vs_Lect, on = .(Lect_vs_Lect)] %>%
  .[, .(Lect, i.Lect, Distance)] %>%
  dcast(Lect ~ i.Lect, value.var = 'Distance') %>%
  .[, !'Lect']

PhonoDistNumbers
```

```
##      A'ou Akajeru Amdo Tibetan Angami Naga Ao Naga Archi Aromanian Arpitan
##   1: 0.00    2.54         5.28         2.80    2.43 5.70     3.77    2.94
##   2: 2.54    0.00         4.51         3.55    3.06 4.58     3.65    1.86
##   3: 5.28    4.51         0.00         3.49    4.27 3.07     6.38    3.50
##   4: 2.80    3.55         3.49         0.00    2.80 4.72     4.39    2.48
##   5: 2.43    3.06         4.27         2.80    0.00 4.33     5.57    3.15
```

```
##  ---
## 331: 4.89  4.68  2.91  2.41  3.40  3.17  7.50  4.28
## 332: 4.69  4.80  2.79  2.71  3.49  3.79  6.86  4.03
## 333: 2.58  3.39  4.03  1.01  2.65  5.06  4.93  2.84
## 334: 5.41  5.50  2.76  3.81  4.69  4.66  6.48  4.40
## 335: 2.13  2.57  3.55  2.42  1.20  3.87  5.22  2.99
## 327 variables not shown: [Arvanitika Albanian, Asho Chin, Assamese, Asturian-Leonese-Cantabrian, Atong (India), Avar, Baba Malay, Badaga, Bagvalal, Bantawa, ...]
```

Create a data.table for clusters.

```
PhonoClusters <-
  Lect_LonLat

PhonoClusters
```

```
##             Lect   lon  lat Dummy
##   1:        A'ou 105.8 26.8 Dummy
##   2:     Akajeru  93.0 13.2 Dummy
##   3: Amdo Tibetan 100.5 34.5 Dummy
##   4:  Angami Naga  93.9 25.7 Dummy
##   5:      Ao Naga  94.4 26.3 Dummy
##  ---
## 331:  Yue Chinese 113.0 23.0 Dummy
## 332:       Zaiwa  98.4 24.2 Dummy
## 333:      Zauzou  98.9 26.5 Dummy
## 334:         Zbu 101.7 32.2 Dummy
## 335:    Zeme Naga  93.6 25.3 Dummy
```

Cluster the lects into two, three, and four groups.

```
PhonoClusters$K2 <-
  PhonoDistNumbers %>%
    kmeans(2) %>%
    pluck(1) %>%
    as_factor()

PhonoClusters$K3 <-
  PhonoDistNumbers %>%
    kmeans(3) %>%
    pluck(1) %>%
    as_factor()

PhonoClusters$K4 <-
  PhonoDistNumbers %>%
    kmeans(4) %>%
    pluck(1) %>%
```

```
  as.numeric() %>%
  as.factor()

PhonoClusters
```

```
##             Lect   lon  lat Dummy K2 K3 K4
##   1:        A'ou 105.8 26.8 Dummy  1  3  2
##   2:     Akajeru  93.0 13.2 Dummy  1  3  2
##   3: Amdo Tibetan 100.5 34.5 Dummy  2  2  4
##   4:  Angami Naga  93.9 25.7 Dummy  1  3  2
##   5:     Ao Naga  94.4 26.3 Dummy  1  3  2
##  ---
## 331:  Yue Chinese 113.0 23.0 Dummy  1  2  1
## 332:       Zaiwa  98.4 24.2 Dummy  2  2  1
## 333:      Zauzou  98.9 26.5 Dummy  1  3  2
## 334:         Zbu 101.7 32.2 Dummy  2  1  3
## 335:    Zeme Naga  93.6 25.3 Dummy  1  3  2
```

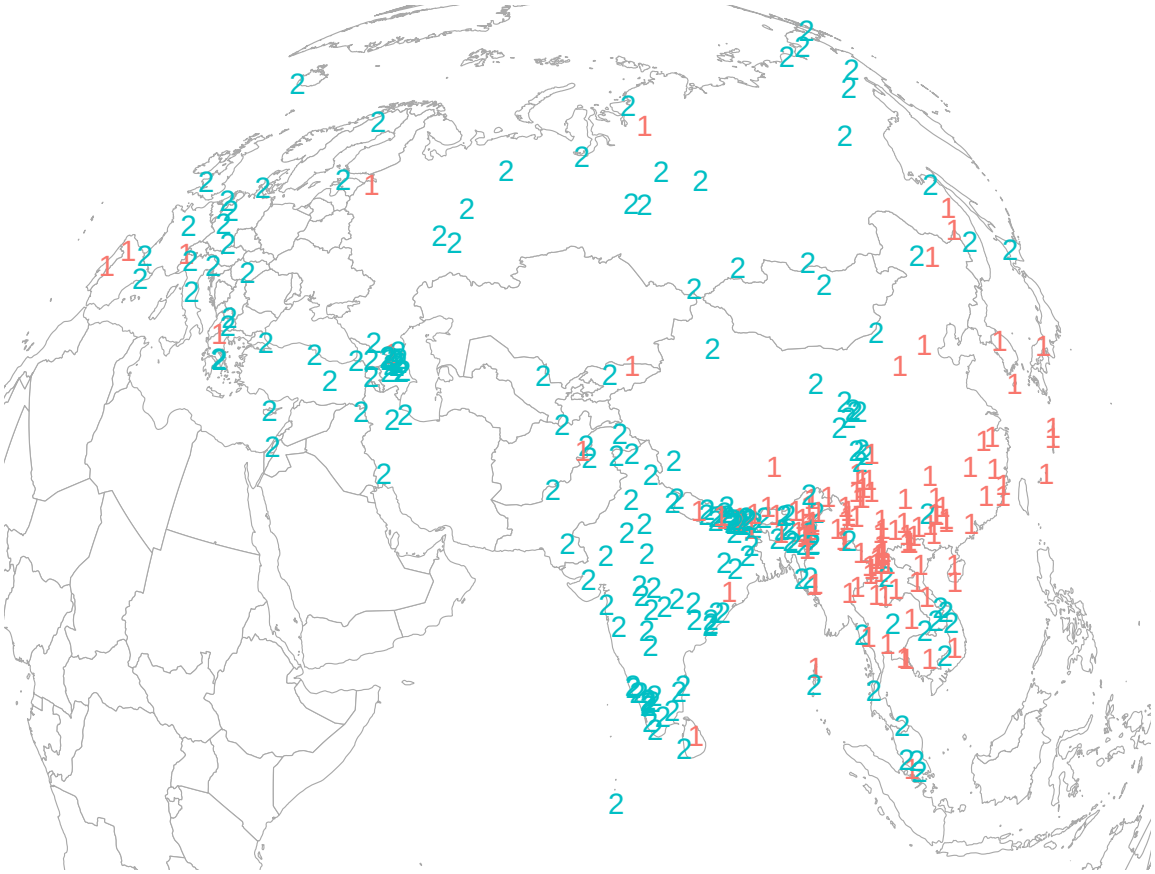
Assign the two clusters on the map of Eurasia, each integer in different colors representing different clusters.

```
PhonoK2 <- EurasiaMap +
 geom_text(aes(x = lon,
               y = lat,
               label = K2,
               color = K2),
           data = PhonoClusters,
           show.legend = FALSE) +
 theme(legend.position = 'bottom')

cairo_pdf(file = "PhonoK2.pdf",
          family = "Times New Roman",
          width = 7,
          height = 5)
 PhonoK2
dev.off()
```

```
## pdf
##   2
```

```
PhonoK2
```



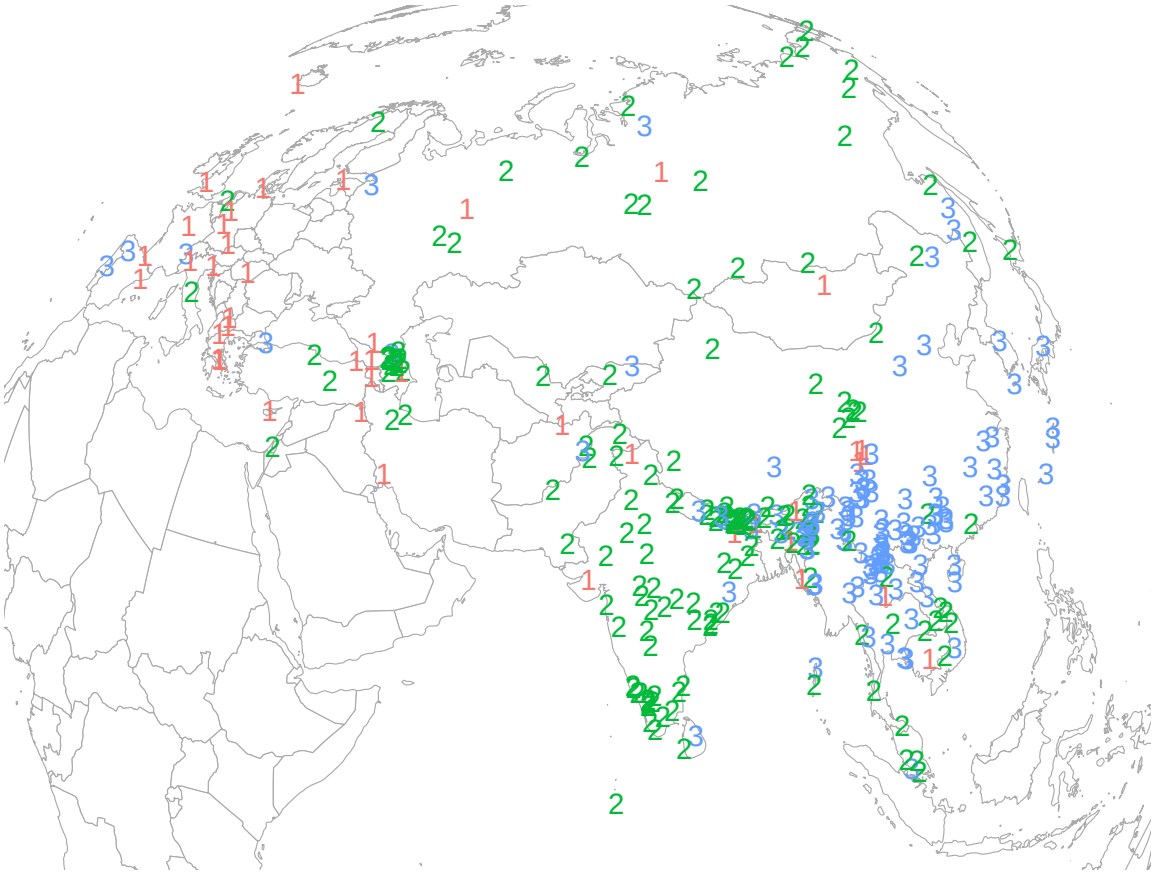
Three clusters on the map.

```
PhonoK3 <- EurasiaMap +
  geom_text(aes(x = lon,
                y = lat,
                label = K3,
               color = K3),
            data = PhonoClusters,
           show.legend = FALSE) +
  theme(legend.position = 'bottom')

cairo_pdf(file = "PhonoK3.pdf",
          family = "Times New Roman",
          width = 7,
          height = 5)
  PhonoK3
dev.off()
```

```
## pdf
##   2
```

```
PhonoK3
```



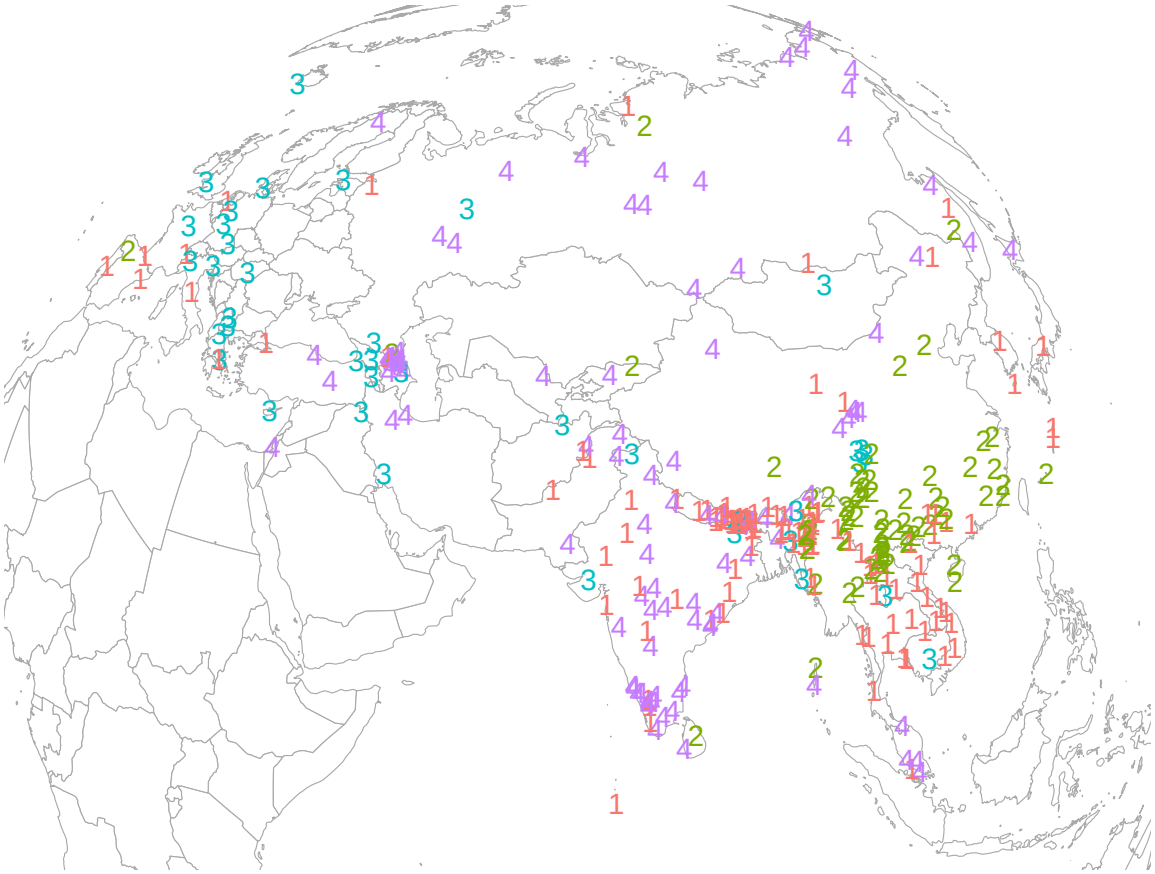
Four clusters on the map.

```
PhonoK4 <- EurasiaMap +
  geom_text(aes(x = lon,
                y = lat,
                label = K4,
                color = K4),
            data = PhonoClusters,
            show.legend = FALSE) +
  theme(legend.position = 'bottom')

cairo_pdf(file = "PhonoK4.pdf",
          family = "Times New Roman",
          width = 7,
          height = 5)
  PhonoK4
dev.off()
```

```
## pdf
##   2
```

```
PhonoK4
```



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.

Subset Coordinates x.

```
Coordinates.x <- select(Lect_vs_Lect, lon, lat)

head(Coordinates.x)
```

```
##      lon  lat
## 1: 105.8 26.8
## 2:  93.0 13.2
## 3: 100.5 34.5
## 4:  93.9 25.7
## 5:  94.4 26.3
## 6:  46.8 42.3
```

Subset Coordinates y.

```
Coordinates.y <- select(Lect_vs_Lect, i.lon, i.lat)

head(Coordinates.y)
```

```
##    i.lon i.lat
## 1:   106  26.8
## 2:   106  26.8
## 3:   106  26.8
## 4:   106  26.8
## 5:   106  26.8
## 6:   106  26.8
```

Calculate the geographical distances between two columns of coordinates. 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.

```
GeoDist <- Lect_vs_Lect %>%
 mutate(Kilometers =
        distHaversine(Coordinates.x, Coordinates.y) / 1000) %>%
 select(Lect_vs_Lect, Kilometers) %>%
 distinct()

GeoDist
```

```
##                 Lect_vs_Lect Kilometers
##     1:          A'ou vs. A'ou          0
##     2:       A'ou vs. Akajeru       2029
##     3:   A'ou vs. Amdo Tibetan       1001
##     4:    A'ou vs. Angami Naga       1202
##     5:       A'ou vs. Ao Naga       1143
##    ---
## 56276:         Zauzou vs. Zbu        691
## 56277:    Zauzou vs. Zeme Naga        537
## 56278:           Zbu vs. Zbu          0
## 56279:      Zbu vs. Zeme Naga       1091
## 56280: Zeme Naga vs. Zeme Naga          0
```

Join the phonological distances to the geographical distances.

```
PhonoGeoDist <- GeoDist %>%
 .[PhonoDist, on = .(Lect_vs_Lect), nomatch = 0] %>%
 .[Lect_vs_Lect, on = .(Lect_vs_Lect), nomatch = 0] %>%
 filter(Lect != i.Lect)

PhonoGeoDist
```

```
##                  Lect_vs_Lect Kilometers Distance        Lect   lon  lat
##      1:         A'ou vs. Akajeru       2029     2.54      Akajeru  93.0 13.2
##      2:     A'ou vs. Amdo Tibetan       1001     5.28 Amdo Tibetan 100.5 34.5
##      3:      A'ou vs. Angami Naga       1202     2.80  Angami Naga  93.9 25.7
##      4:         A'ou vs. Ao Naga       1143     2.43      Ao Naga  94.4 26.3
##      5:           A'ou vs. Archi       5562     5.70        Archi  46.8 42.3
##     ---
## 111886: Youle Jinuo vs. Zeme Naga        840     2.30  Youle Jinuo 101.1 22.0
## 111887: Yue Chinese vs. Zeme Naga       1981     2.62  Yue Chinese 113.0 23.0
## 111888:       Zaiwa vs. Zeme Naga        495     2.92        Zaiwa  98.4 24.2
## 111889:      Zauzou vs. Zeme Naga        537     2.48       Zauzou  98.9 26.5
## 111890:         Zbu vs. Zeme Naga       1091     4.12          Zbu 101.7 32.2
## 3 variables not shown: [i.Lect, i.lon, i.lat]
```

Set Burushaski, a language isolate spoken in the middle of Eurasia, as a reference point. The hypothesis is that, the closer a Eurasian Lect is to Burushaski geographically, the closer it is to Burushaski phonologically.

```
Burushaski <- PhonoGeoDist %>%
 filter(i.Lect == 'Burushaski') %>%
 select(-i.Lect, -i.lon, -i.lat)

Burushaski
```

```
##                  Lect_vs_Lect Kilometers Distance        Lect   lon  lat
##   1:         A'ou vs. Burushaski       3109     5.41         A'ou 105.8 26.8
##   2:      Akajeru vs. Burushaski       3142     4.28      Akajeru  93.0 13.2
##   3: Amdo Tibetan vs. Burushaski       2330     2.35 Amdo Tibetan 100.5 34.5
##   4:  Angami Naga vs. Burushaski       2157     4.21  Angami Naga  93.9 25.7
##   5:      Ao Naga vs. Burushaski       2156     4.11      Ao Naga  94.4 26.3
##  ---
## 330:  Burushaski vs. Yue Chinese       3946     2.81  Yue Chinese 113.0 23.0
## 331:       Burushaski vs. Zaiwa       2621     3.09        Zaiwa  98.4 24.2
## 332:      Burushaski vs. Zauzou       2521     4.66       Zauzou  98.9 26.5
## 333:         Burushaski vs. Zbu       2508     3.78          Zbu 101.7 32.2
## 334:    Burushaski vs. Zeme Naga       2162     3.51    Zeme Naga  93.6 25.3
```

Subset the coordinates of the Burushaski data.table.

```
BurushaskiCoords <-
 Burushaski %>%
 select(lon, lat)

BurushaskiCoords
```

```
##       lon  lat
##   1: 105.8 26.8
```

```
##   2:  93.0 13.2
##   3: 100.5 34.5
##   4:  93.9 25.7
##   5:  94.4 26.3
##  ---
## 330: 113.0 23.0
## 331:  98.4 24.2
## 332:  98.9 26.5
## 333: 101.7 32.2
## 334:  93.6 25.3
```

Among all the lects other than Burushaski, calculate the greatest distance from a lect to its nearest neighbor.

```
MaxDistance <-
  BurushaskiCoords %>%
  knearneigh(k = 1, longlat = TRUE) %>%
  knn2nb() %>%
  nbdists(BurushaskiCoords, longlat = TRUE) %>%
  unlist() %>%
  max()

MaxDistance
```

```
## [1] 1552
```

Assign the neighbors to each lect, a neighbor defined as a lect within the maximal distance calculated above, so that every lect has at least one neighbor.

```
WeightMatrix <-
  BurushaskiCoords %>%
  dnearneigh(0, MaxDistance, longlat = TRUE) %>%
  nb2listw(style = 'B')

WeightMatrix
```

```
## Characteristics of weights list object:
## Neighbour list object:
## Number of regions: 334
## Number of nonzero links: 28440
## Percentage nonzero weights: 25.5
## Average number of links: 85.1
##
## Weights style: B
## Weights constants summary:
##     n     nn    S0    S1       S2
## B 334 111556 28440 56880 13949472
```

Perform the Moran's I test to test the spatial autocorrelation, which shows that the phonological distance patterns are areally clustered.

```
moran.test(Burushaski$Distance, WeightMatrix)
```

```
##
##  Moran I test under randomisation
##
## data:  Burushaski$Distance
## weights: WeightMatrix
##
## Moran I statistic standard deviate = 46, p-value <0.0000000000000002
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##         0.2745861        -0.0030030         0.0000366
```

Perform a spatial regression analysis based on the spatial lag model. The results show that geographical distance to Burushaski is positively correlated with the phonological distance to it.

```
SpatialLag <-
  lagsarlm(Distance ~ Kilometers,
        Burushaski,
        listw = WeightMatrix)

summary(SpatialLag)
```

```
##
## Call:lagsarlm(formula = Distance ~ Kilometers, data = Burushaski,
##     listw = WeightMatrix)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16356 -0.70537 -0.13814  0.60314  4.45891
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value          Pr(>|z|)
## (Intercept) 1.00723029 0.18620702  5.4092 0.0000000633083141
## Kilometers  0.00031947 0.00004472  7.1438 0.0000000000009077
##
## Rho: 0.00347, LR test value: 116, p-value: < 0.000000000000000222
## Asymptotic standard error: 0.000297
##     z-value: 11.7, p-value: < 0.000000000000000222
## Wald statistic: 137, p-value: < 0.000000000000000222
##
## Log likelihood: -469 for lag model
```

```
## ML residual variance (sigma squared): 0.968, (sigma: 0.984)
## Number of observations: 334
## Number of parameters estimated: 4
## AIC: 946, (AIC for lm: 1060)
## LM test for residual autocorrelation
## test value: 16.1, p-value: 0.000061165
```

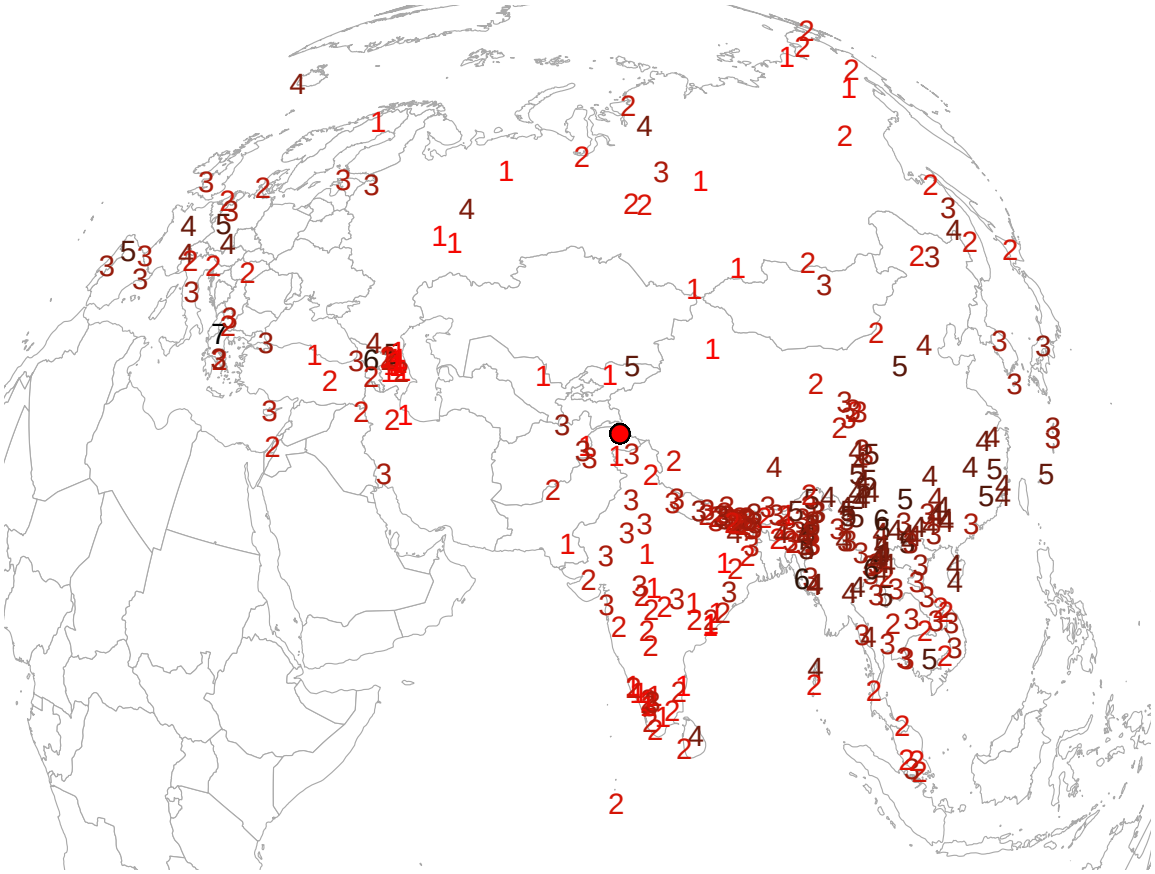
Visualize the phonological distances to Burushaski on a map of Eurasia.

```
BurushaskiMap <- EurasiaMap +
  geom_text(aes(x = lon,
                y = lat,
                color = Distance,
                label = as.integer(round(Distance))),
            data = Burushaski,
           show.legend = FALSE) +
  geom_point(x = 74.8, y = 36.2, color = 'black', fill = 'red', shape = 21, size = 3) +
  scale_color_gradient(low = 'red', high = 'black')

cairo_pdf(file = "BurushaskiMap.pdf",
          family = "Times New Roman",
          width = 7,
          height = 5)
  BurushaskiMap
dev.off()
```

```
## pdf
##   2
```

```
BurushaskiMap
```



We see that generally, geographically closer lects to Burushaski (red dot) also tend to be phonologically closer to it.

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?

Divide the Eurasian lects into seven different regions solely based on their geographical coordinates: Northeast Asia, Mainland Southeast Asia, Qinghai-Gansu, South Asia, West Asia, Caucasus, and Europe.

```
Areas <- Lect_LonLat %>%
  mutate(Area =
         ifelse(lon > 90 & lon < 105 & lat > 32 & lat < 40,
                'Qinghai-Gansu',
         ifelse(lon > 90 & lat <= 32,
                'Mainland Southeast Asia',
         ifelse(lon > 60 & lon <= 90 & lat < 40,
                'South Asia',
         ifelse(lon > 25 & lon < 50 & lat < 50,
                'West Asia',
         ifelse(lon > -25 & lon < 50 & lat > 30,
```

```
              'Europe',
              'Northeast Asia'))))))) %>%
  select(-lon, -lat)

Areas
```

```
##             Lect Dummy                  Area
##   1:         A'ou Dummy Mainland Southeast Asia
##   2:      Akajeru Dummy Mainland Southeast Asia
##   3: Amdo Tibetan Dummy          Qinghai-Gansu
##   4:  Angami Naga Dummy Mainland Southeast Asia
##   5:       Ao Naga Dummy Mainland Southeast Asia
##  ---
## 331:  Yue Chinese Dummy Mainland Southeast Asia
## 332:        Zaiwa Dummy Mainland Southeast Asia
## 333:       Zauzou Dummy Mainland Southeast Asia
## 334:          Zbu Dummy          Qinghai-Gansu
## 335:    Zeme Naga Dummy Mainland Southeast Asia
```

The map visualizes the lects in the predefined seven areas.

```
AreasLonLat <- Areas %>%
  left_join(Lect_LonLat)
```

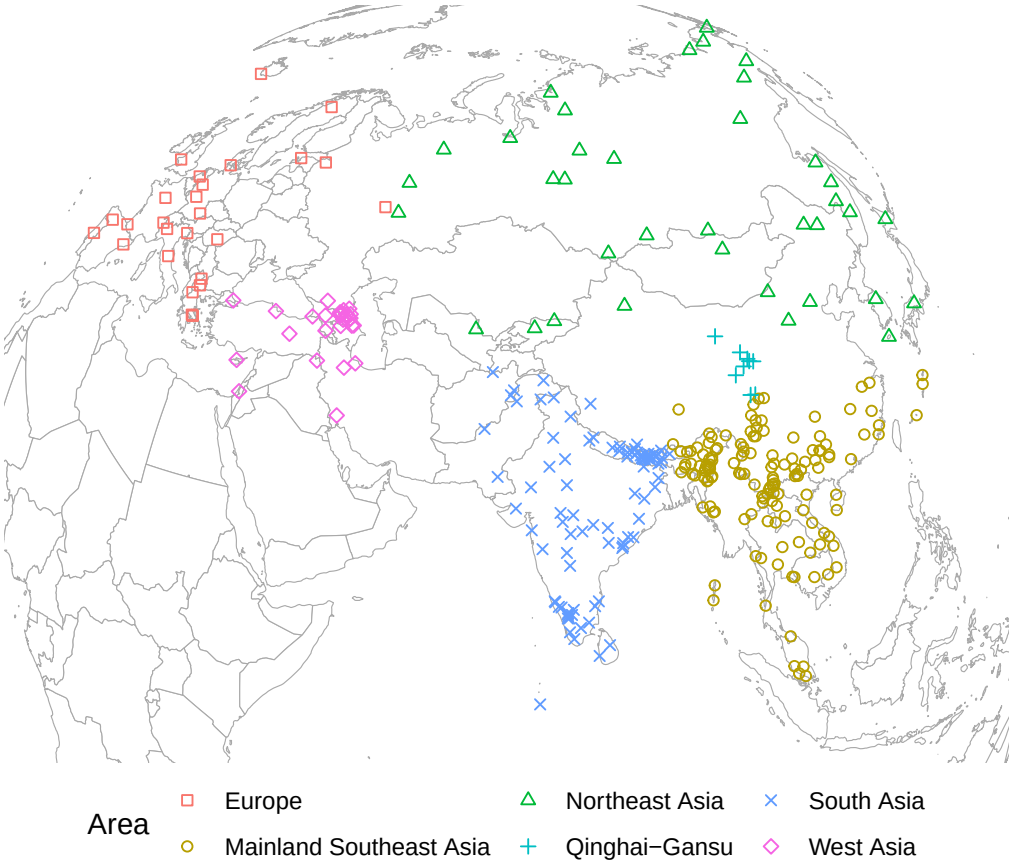
```
## Joining with `by = join_by(Lect, Dummy)`
```

```
AreaMap <- EurasiaMap +
  geom_point(aes(x = lon,
                 y = lat,
                 color = Area,
                 shape = Area),
             data = AreasLonLat) +
  scale_shape_manual(values = 0:5) +
  theme(legend.position = 'bottom')

cairo_pdf(file = "AreaMap.pdf",
          family = "Times New Roman",
          width = 6,
          height = 5)
  AreaMap
dev.off()
```

```
## pdf
##   2
```

```
AreaMap
```



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

Split the Lect vs. Lect column of the PhonoDist table into two lects.

```
Lect_iLect <- str_split_fixed(
  PhonoDist$Lect_vs_Lect, ' vs. ', n = 2) %>%
  as.data.table() %>%
  setnames(c('Lect', 'i.Lect'))

Lect_iLect
```

```
##           Lect       i.Lect
##     1:    A'ou         A'ou
##     2:    A'ou      Akajeru
##     3:    A'ou Amdo Tibetan
##     4:    A'ou  Angami Naga
##     5:    A'ou      Ao Naga
##    ---
## 56276:  Zauzou          Zbu
## 56277:  Zauzou    Zeme Naga
## 56278:     Zbu          Zbu
## 56279:     Zbu    Zeme Naga
## 56280: Zeme Naga    Zeme Naga
```

Join the distances to the split lect pairs.

```
PhonoDistSplit <- PhonoDist %>%
  bind_cols(Lect_iLect) %>%
  select(-Lect_vs_Lect)

PhonoDistSplit
```

```
##        Distance      Lect       i.Lect
##     1:     0.00      A'ou        A'ou
##     2:     2.54      A'ou      Akajeru
##     3:     5.28      A'ou Amdo Tibetan
##     4:     2.80      A'ou  Angami Naga
##     5:     2.43      A'ou      Ao Naga
##    ---
## 56276:     4.68    Zauzou         Zbu
## 56277:     2.48    Zauzou    Zeme Naga
## 56278:     0.00       Zbu         Zbu
## 56279:     4.12       Zbu    Zeme Naga
## 56280:     0.00 Zeme Naga    Zeme Naga
```

Double the split distances by switching Lect.x and Lect.y.

```
PhonoDistDouble <- PhonoDistSplit %>%
  rename(c('Lect' = 'i.Lect',
           'i.Lect' = 'Lect')) %>%
  bind_rows(PhonoDistSplit)

PhonoDistDouble
```

```
##        Distance    i.Lect        Lect
##     1:     0.00      A'ou        A'ou
##     2:     2.54      A'ou      Akajeru
##     3:     5.28      A'ou Amdo Tibetan
##     4:     2.80      A'ou  Angami Naga
##     5:     2.43      A'ou      Ao Naga
##    ---
## 112556:     4.68       Zbu       Zauzou
## 112557:     2.48 Zeme Naga       Zauzou
## 112558:     0.00       Zbu         Zbu
## 112559:     4.12 Zeme Naga         Zbu
## 112560:     0.00 Zeme Naga    Zeme Naga
```

Make a matrix consisting of lects, their distances from all other lects, and their area.

```
PhonoDistWide <- PhonoDistDouble %>%
  pivot_wider(names_from = i.Lect,
              values_from = Distance,
              values_fn = sum) %>%
  left_join(Areas) %>%
  as.data.table()
```

```
## Joining with `by = join_by(Lect)`
```

```
setcolorder(PhonoDistWide,
            c(ncol(PhonoDistWide),
            1:(ncol(PhonoDistWide) - 1)))

PhonoDistWide
```

```
##                    Area         Lect A'ou Akajeru Amdo Tibetan Angami Naga
##   1: Mainland Southeast Asia          A'ou 0.00    2.54         5.28        2.80
##   2: Mainland Southeast Asia       Akajeru 2.54    0.00         4.51        3.55
##   3:           Qinghai-Gansu Amdo Tibetan 5.28    4.51         0.00        3.49
##   4: Mainland Southeast Asia  Angami Naga 2.80    3.55         3.49        0.00
##   5: Mainland Southeast Asia       Ao Naga 2.43    3.06         4.27        2.80
##  ---
## 331: Mainland Southeast Asia  Yue Chinese 4.89    4.68         2.91        2.41
## 332: Mainland Southeast Asia        Zaiwa 4.69    4.80         2.79        2.71
## 333: Mainland Southeast Asia       Zauzou 2.58    3.39         4.03        1.01
## 334:           Qinghai-Gansu          Zbu 5.41    5.50         2.76        3.81
## 335: Mainland Southeast Asia    Zeme Naga 2.13    2.57         3.55        2.42
## 332 variables not shown: [Ao Naga, Archi, Aromanian, Arpitan, Arvanitika Albanian, Asho Chin, Assamese, Asturian
Leonese-Cantabrian, Atong (India), Avar, ...]
```

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

```
AreaSample <- Areas %>%
  group_by(Area) %>%
  slice_sample(prop = 0.5)

AreaSample
```

```
## # A tibble: 166 x 3
## # Groups:   Area [6]
##    Lect           Dummy Area
##    <chr>          <chr> <chr>
##  1 Danish         Dummy Europe
##  2 Modern Greek   Dummy Europe
```

```
##  3 Gheg Albanian    Dummy Europe
##  4 Hungarian        Dummy Europe
##  5 Estonian Swedish Dummy Europe
##  6 Friulian         Dummy Europe
##  7 French           Dummy Europe
##  8 Macedonian       Dummy Europe
##  9 Aromanian        Dummy Europe
## 10 Italian          Dummy Europe
## # i 156 more rows
```

Subset the first half of the lects and their distances, which I will train the Naive Bayes Classifier with.

```
Train <- PhonoDistWide[Lect %in% AreaSample$Lect]

Train
```

```
##                    Area                 Lect A'ou Akajeru Amdo Tibetan
##   1:          Qinghai-Gansu          Amdo Tibetan 5.28    4.51         0.00
##   2: Mainland Southeast Asia           Angami Naga 2.80    3.55         3.49
##   3:              West Asia                Archi 5.70    4.58         3.07
##   4:                  Europe            Aromanian 3.77    3.65         6.38
##   5: Mainland Southeast Asia             Asho Chin 5.01    5.02         2.68
##  ---
## 162:                  Europe           Westphalic 5.21    4.17         4.17
## 163: Mainland Southeast Asia           Wu Chinese 1.85    2.72         3.79
## 164: Mainland Southeast Asia Yerong-Southern Buyang 2.07    2.68         3.42
## 165: Mainland Southeast Asia                Zauzou 2.58    3.39         4.03
## 166: Mainland Southeast Asia             Zeme Naga 2.13    2.57         3.55
## 333 variables not shown: [Angami Naga, Ao Naga, Archi, Aromanian, Arpitan, Arvanitika Albanian, Asho Chin, Assa
Leonese-Cantabrian, Atong (India), ...]
```

The remaining half will be the lects where the Naive Bayes Classifier will be tested upon to predict their areas.

```
Test <- PhonoDistWide %>%
 filter(!(Lect %in% Train$Lect)) %>%
 select(-Area)

Test
```

```
##                 Lect A'ou Akajeru Amdo Tibetan Angami Naga Ao Naga Archi
##   1:            A'ou 0.00    2.54         5.28        2.80    2.43  5.70
##   2:          Akajeru 2.54    0.00         4.51        3.55    3.06  4.58
##   3:          Ao Naga 2.43    3.06         4.27        2.80    0.00  4.33
##   4:          Arpitan 2.94    1.86         3.50        2.48    3.15  4.29
##   5: Arvanitika Albanian 4.32    3.18         2.64        4.67    4.49  3.48
```

```
##  ---
## 165:      Yongbei Zhuang 2.24    2.72        3.31        2.38    1.71  3.89
## 166:        Youle Jinuo 2.23    3.18        3.44        1.99    2.29  4.54
## 167:        Yue Chinese 4.89    4.68        2.91        2.41    3.40  3.17
## 168:              Zaiwa 4.69    4.80        2.79        2.71    3.49  3.79
## 169:                Zbu 5.41    5.50        2.76        3.81    4.69  4.66
## 330 variables not shown: [Aromanian, Arpitan, Arvanitika Albanian, Asho Chin, Assamese, Asturian-
Leonese-Cantabrian, Atong (India), Avar, Baba Malay, Badaga, ...]
```

Train the classifier.

```
Classifier <- naiveBayes(Area ~ ., Train)

Classifier[1]
```

```
## $apriori
## Y
##                Europe Mainland Southeast Asia         Northeast Asia
##                    13                    70                    18
##         Qinghai-Gansu            South Asia            West Asia
##                     4                    47                    14
```

The trained classifier then predicts the areas of the remaining lects.

```
Predict <- predict(Classifier, newdata = Test)

Predict
```

```
##   [1] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
##   [4] Mainland Southeast Asia Europe                  South Asia
##   [7] Europe                  South Asia              Northeast Asia
##  [10] West Asia               South Asia              South Asia
##  [13] South Asia              Mainland Southeast Asia Mainland Southeast Asia
##  [16] Northeast Asia          West Asia               Mainland Southeast Asia
##  [19] South Asia              Mainland Southeast Asia Europe
##  [22] Northeast Asia          Mainland Southeast Asia Northeast Asia
##  [25] Northeast Asia          Europe                  Mainland Southeast Asia
##  [28] South Asia              West Asia               South Asia
##  [31] South Asia              Qinghai-Gansu           Qinghai-Gansu
##  [34] South Asia              Mainland Southeast Asia South Asia
##  [37] Mainland Southeast Asia South Asia              Mainland Southeast Asia
##  [40] Europe                  South Asia              South Asia
##  [43] Mainland Southeast Asia Mainland Southeast Asia South Asia
##  [46] Europe                  South Asia              Europe
##  [49] South Asia              Europe                  West Asia
##  [52] Mainland Southeast Asia Northeast Asia          Europe
##  [55] South Asia              Northeast Asia          Europe
```

```
##  [58] Northeast Asia         Mainland Southeast Asia Mainland Southeast Asia
##  [61] Europe                 Northeast Asia          South Asia
##  [64] Europe                 Qinghai-Gansu           Mainland Southeast Asia
##  [67] South Asia             Mainland Southeast Asia Northeast Asia
##  [70] Mainland Southeast Asia Mainland Southeast Asia South Asia
##  [73] Northeast Asia         Mainland Southeast Asia South Asia
##  [76] Northeast Asia         South Asia              South Asia
##  [79] South Asia             Mainland Southeast Asia South Asia
##  [82] South Asia             Mainland Southeast Asia South Asia
##  [85] Mainland Southeast Asia South Asia             Mainland Southeast Asia
##  [88] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
##  [91] Northeast Asia         Mainland Southeast Asia Mainland Southeast Asia
##  [94] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
##  [97] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
## [100] Europe                 Northeast Asia          South Asia
## [103] Mainland Southeast Asia Mainland Southeast Asia Northeast Asia
## [106] Mainland Southeast Asia Northeast Asia         South Asia
## [109] South Asia             Mainland Southeast Asia West Asia
## [112] West Asia              Mainland Southeast Asia South Asia
## [115] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
## [118] Northeast Asia         Mainland Southeast Asia Mainland Southeast Asia
## [121] South Asia             Mainland Southeast Asia Mainland Southeast Asia
## [124] Northeast Asia         Europe                  South Asia
## [127] South Asia             South Asia              Europe
## [130] Europe                 Mainland Southeast Asia Northeast Asia
## [133] Qinghai-Gansu          South Asia              Northeast Asia
## [136] Northeast Asia         Northeast Asia          Mainland Southeast Asia
## [139] Northeast Asia         Northeast Asia          Northeast Asia
## [142] South Asia             Mainland Southeast Asia Northeast Asia
## [145] Qinghai-Gansu          Northeast Asia          Mainland Southeast Asia
## [148] Mainland Southeast Asia South Asia             Mainland Southeast Asia
## [151] South Asia             Qinghai-Gansu           Northeast Asia
## [154] Northeast Asia         Northeast Asia          South Asia
## [157] South Asia             South Asia              Europe
## [160] Qinghai-Gansu          Mainland Southeast Asia Mainland Southeast Asia
## [163] Qinghai-Gansu          Qinghai-Gansu           Mainland Southeast Asia
## [166] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
## [169] Europe
## 6 Levels: Europe Mainland Southeast Asia Northeast Asia ... West Asia
```

Join the predicted areas into the table of tested lects.

```
AreaPrediction1 <- Test %>%
  select(Lect) %>%
  mutate(Prediction = Predict) %>%
  left_join(Areas) %>%
  mutate(Correct = Prediction == Area)
```

```
## Joining with `by = join_by(Lect)`
```

```
AreaPrediction1
```

```
##                  Lect            Prediction Dummy                  Area
##   1:             A'ou Mainland Southeast Asia Dummy Mainland Southeast Asia
##   2:          Akajeru Mainland Southeast Asia Dummy Mainland Southeast Asia
##   3:          Ao Naga Mainland Southeast Asia Dummy Mainland Southeast Asia
##   4:          Arpitan Mainland Southeast Asia Dummy                Europe
##   5: Arvanitika Albanian                Europe Dummy                Europe
##  ---
## 165:     Yongbei Zhuang Mainland Southeast Asia Dummy Mainland Southeast Asia
## 166:       Youle Jinuo Mainland Southeast Asia Dummy Mainland Southeast Asia
## 167:       Yue Chinese Mainland Southeast Asia Dummy Mainland Southeast Asia
## 168:            Zaiwa Mainland Southeast Asia Dummy Mainland Southeast Asia
## 169:              Zbu                Europe Dummy         Qinghai-Gansu
## 1 variable not shown: [Correct]
```

Perform the same training and testing, but training with the latter half as the training and testing the former half.

```
Train2 <- Test %>%
  left_join(Areas)
```

```
## Joining with `by = join_by(Lect, Dummy)`
```

```
Test2 <- Train %>%
  select(-Area)

Classifier <- naiveBayes(Area ~ ., Train2)

Predict2 <- predict(Classifier, newdata = Test2)

AreaPrediction2 <- Test2 %>%
  select(Lect) %>%
  mutate(Prediction = Predict2) %>%
  left_join(Areas) %>%
  mutate(Correct = Prediction == Area)
```

```
## Joining with `by = join_by(Lect)`
```

```
AreaPrediction2
```

```
##                  Lect            Prediction Dummy
##   1:       Amdo Tibetan         Qinghai-Gansu Dummy
##   2:        Angami Naga Mainland Southeast Asia Dummy
##   3:              Archi            West Asia Dummy
##   4:          Aromanian                Europe Dummy
```

```
##   5:            Asho Chin            South Asia Dummy
##  ---
## 162:           Westphalic                Europe Dummy
## 163:           Wu Chinese Mainland Southeast Asia Dummy
## 164: Yerong-Southern Buyang Mainland Southeast Asia Dummy
## 165:              Zauzou Mainland Southeast Asia Dummy
## 166:           Zeme Naga Mainland Southeast Asia Dummy
## 2 variables not shown: [Area, Correct]
```

Join the two halves whose areas are predicted based on each other.

```
AreaPrediction <- bind_rows(AreaPrediction1, AreaPrediction2)

AreaPrediction
```

```
##                   Lect            Prediction Dummy
##   1:              A'ou Mainland Southeast Asia Dummy
##   2:           Akajeru Mainland Southeast Asia Dummy
##   3:           Ao Naga Mainland Southeast Asia Dummy
##   4:           Arpitan Mainland Southeast Asia Dummy
##   5:  Arvanitika Albanian                Europe Dummy
##  ---
## 331:           Westphalic                Europe Dummy
## 332:           Wu Chinese Mainland Southeast Asia Dummy
## 333: Yerong-Southern Buyang Mainland Southeast Asia Dummy
## 334:              Zauzou Mainland Southeast Asia Dummy
## 335:           Zeme Naga Mainland Southeast Asia Dummy
## 2 variables not shown: [Area, Correct]
```

I will analyze how correctly the model has predicted the areas based on confusion matrix.

Make the predefined areas into factors.

```
AreaFactors <- as.factor(AreaPrediction$Area)

head(AreaFactors)
```

```
## [1] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
## [4] Europe                  Europe                  Mainland Southeast Asia
## 6 Levels: Europe Mainland Southeast Asia Northeast Asia ... West Asia
```

Make the predicted areas into factors.

```
PredictionFactors <- factor(AreaPrediction$Prediction,
                            levels = levels(AreaFactors))

head(PredictionFactors)
```

```
## [1] Mainland Southeast Asia Mainland Southeast Asia Mainland Southeast Asia
## [4] Mainland Southeast Asia Europe                  South Asia
## 6 Levels: Europe Mainland Southeast Asia Northeast Asia ... West Asia
```

Below is the confusion matrix and the related statistics, based on the combination of the two halves of prediction:

```
ConfusionMatrix <-
  confusionMatrix(PredictionFactors,
                  AreaFactors,
                  mode = 'everything')

ConfusionMatrixOverall <-
  ConfusionMatrix$overall %>%
  as.matrix() %>%
  as.data.frame() %>%
  rownames_to_column()

colnames(ConfusionMatrixOverall) <- c('Name', 'Value')

ConfusionMatrixOverall %>%
    xtable(label = 'ConfusionMatrix',
           caption = 'Confusion matrix based on two halves of Naive Bayes Classifier prediction',
           style = 'latex') %>%
    booktabs('ConfusionMatrix.tex')

ConfusionMatrixOverall
```

```
##              Name              Value
## 1        Accuracy 0.63880597014925378
## 2           Kappa 0.51280633646230223
## 3  AccuracyLower 0.58482659468676812
## 4  AccuracyUpper 0.69030683715591790
## 5   AccuracyNull 0.42089552238805972
## 6 AccuracyPValue 0.00000000000000082
## 7  McnemarPValue 0.00617919702119709
```

The F1 values of individual classes.

```
F1 <- ConfusionMatrix$byClass %>%
  as.data.frame() %>%
  rownames_to_column() %>%
  rename(c('Class' = 'rowname')) %>%
  mutate(Class = gsub('Class: ', '', Class)) %>%
  select(Class, F1)

F1 %>%
```

```
 xtable(label = 'F1',
        caption = 'F1 values of individual classes',
        style = 'latex') %>%
 booktabs('F1.tex')

F1
```

```
##                  Class    F1
## 1                Europe 0.476
## 2 Mainland Southeast Asia 0.799
## 3         Northeast Asia 0.553
## 4          Qinghai-Gansu 0.381
## 5            South Asia 0.573
## 6             West Asia 0.491
```

The visualization of the lects by their predicted areas.

```
AreaPredictionLonLat <- AreaPrediction %>%
 left_join(Lect_LonLat)
```

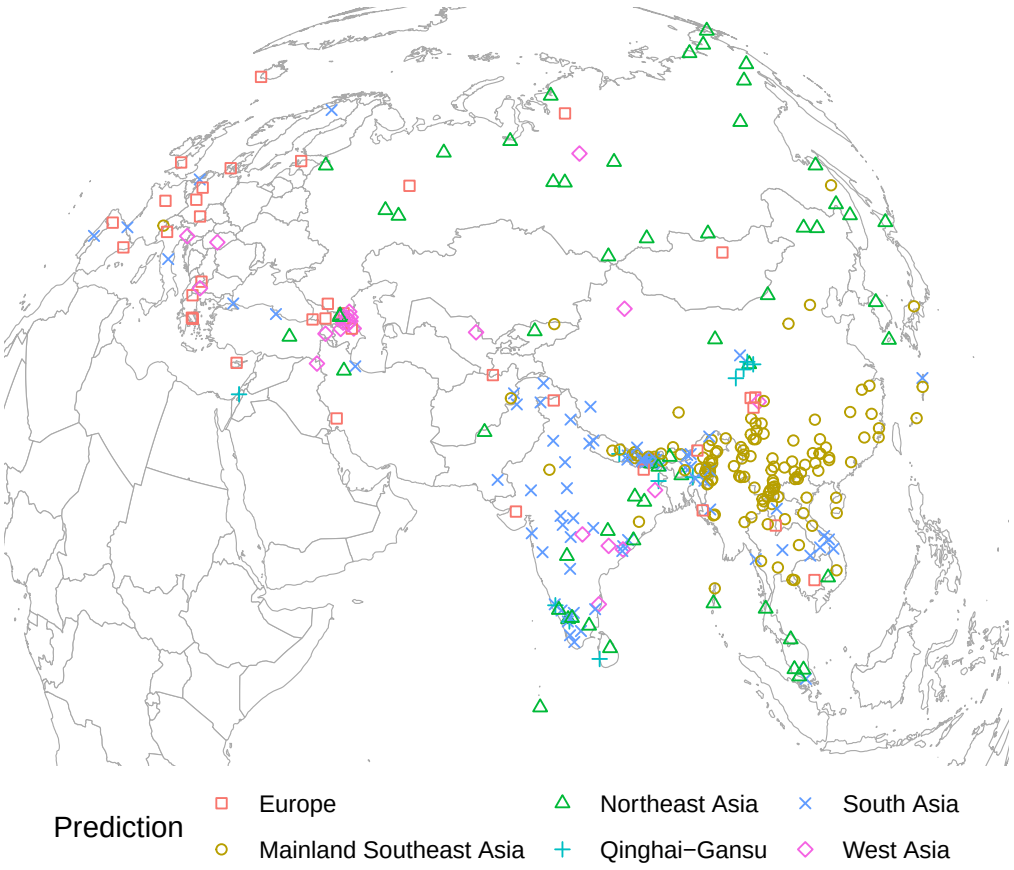
```
## Joining with `by = join_by(Lect, Dummy)`
```

```
AreaPredictionMap <- EurasiaMap +
 geom_point(aes(x = lon,
                y = lat,
                color = Prediction,
                shape = Prediction),
            data = AreaPredictionLonLat) +
 scale_shape_manual(values = 0:6) +
 theme(legend.position = 'bottom')

cairo_pdf(file = "AreaPredictionMap.pdf",
          family = "Times New Roman",
          width = 6,
          height = 5)
 AreaPredictionMap
dev.off()
```

```
## pdf
##   2
```

```
AreaPredictionMap
```



Grambank

Load Eurasian lects of Grambank.

```
GrambankLects <- fread('GrambankLects.csv') %>%
 as.data.table() %>%
 .[Macroarea == 'Eurasia'] %>%
 .[, .(ID, Longitude, Latitude)]

GrambankLects
```

```
##           ID Longitude Latitude
##   1: abkh1244      41.2     43.1
##   2: acha1249      97.7     24.3
##   3: aghw1237      47.1     40.4
##   4: aheu1239     104.2     17.8
##   5: ahom1240      88.5     27.3
##  ---
## 559: zaiw1241      98.4     24.2
## 560: zakh1243      97.0     28.1
## 561: zaye1235      96.9     20.2
## 562: zeme1240      93.6     25.3
## 563: zhab1238     101.0     30.7
```

Load Grambank and subset to lects present in PanPhon.

```
Grambank <- fread('Grambank.csv') %>%
  as.data.table() %>%
  .[Language_ID %in% GrambankLects$ID] %>%
  .[, .(Language_ID, Parameter_ID, Value)]

Grambank
```

```
##        Language_ID Parameter_ID Value
##     1:    abkh1244       GB020     1
##     2:    abkh1244       GB021     1
##     3:    abkh1244       GB022     1
##     4:    abkh1244       GB023     1
##     5:    abkh1244       GB024     3
##    ---
## 106332:    zhab1238       GB433     0
## 106333:    zhab1238       GB519     1
## 106334:    zhab1238       GB520     1
## 106335:    zhab1238       GB521     1
## 106336:    zhab1238       GB522     ?
```

Detect non-binary features.

```
NonBinary <- Grambank %>%
  .[Value == 2] %>%
  .[, .(Parameter_ID)] %>%
  distinct() %>%
  unlist() %>%
  as.vector()

NonBinary
```

```
## [1] "GB203" "GB193" "GB024" "GB025" "GB065" "GB130"
```

Exclude non-binary features and numeralize the values.

```
GrambankBinary <- Grambank %>%
  .[!(Parameter_ID %in% NonBinary)] %>%
  .[, Value := gsub(0, -1, Value)] %>%
  .[, Value := gsub('\\?', 0, Value)] %>%
  .[, Value := as.numeric(Value)]

GrambankBinary
```

Calculate the proportion of unknown values.

```
Unknown <-  GrambankBinary %>%
  .[Value == 0] %>%
  .[, .N]/nrow(GrambankBinary)

Unknown
```

```
## [1] 0.131
```

Create a vector of glottocodes.

```
Glottocodes <- GrambankBinary$Language_ID %>%
  unique()

Glottocodes[1:10]
```

```
##  [1] "abkh1244" "acha1249" "aghw1237" "aheu1239" "ahom1240" "ainu1240"
##  [7] "aito1238" "akab1249" "akaj1239" "akha1245"
```

Widen the Grambank data table.

```
GramDist <- GrambankBinary %>%
  dcast(Language_ID ~ Parameter_ID, value.var = 'Value', fill = 0) %>%
  .[, Language_ID := NULL] %>%
  dist(method = 'manhattan')

GramDist[1:10]
```

```
##  [1] 153 149 148 156 122 180 122 169 151 148
```

K-means clustering (2).

```
GramK2 <- GramDist %>%
  kmeans(2) %>%
  pluck(1) %>%
  as_factor()

GramK2[1:10]
```

```
##  1  2  3  4  5  6  7  8  9 10
##  1  2  1  2  2  1  2  1  1  2
## Levels: 1 2
```

K-means clustering (3).

```
GramK3 <- GramDist %>%
 kmeans(3) %>%
 pluck(1) %>%
 as_factor()

GramK3[1:10]
```

```
##  1  2  3  4  5  6  7  8  9 10
##  2  1  2  1  1  2  2  3  2  1
## Levels: 1 2 3
```

K-means clustering (4).

```
GramK4 <- GramDist %>%
 kmeans(4) %>%
 pluck(1) %>%
 as_factor()

GramK4[1:10]
```

```
##  1  2  3  4  5  6  7  8  9 10
##  3  1  3  1  1  2  2  4  2  1
## Levels: 1 2 3 4
```

Make a data.table of Glottocodes, coordinates, and K

```
GramK <-
 data.table(Glottocode = Glottocodes,
            K2 = GramK2,
            K3 = GramK3,
            K4 = GramK4) %>%
 .[GrambankLects, on = .(Glottocode = ID), nomatch = 0]

GramK
```

```
##      Glottocode K2 K3 K4 Longitude Latitude
##   1:   abkh1244  1  2  3      41.2     43.1
##   2:   acha1249  2  1  1      97.7     24.3
##   3:   aghw1237  1  2  3      47.1     40.4
##   4:   aheu1239  2  1  1     104.2     17.8
##   5:   ahom1240  2  1  1      88.5     27.3
##  ---
## 559:   zaiw1241  2  1  1      98.4     24.2
## 560:   zakh1243  1  3  4      97.0     28.1
## 561:   zaye1235  2  1  1      96.9     20.2
## 562:   zeme1240  1  3  4      93.6     25.3
## 563:   zhab1238  2  1  1     101.0     30.7
```

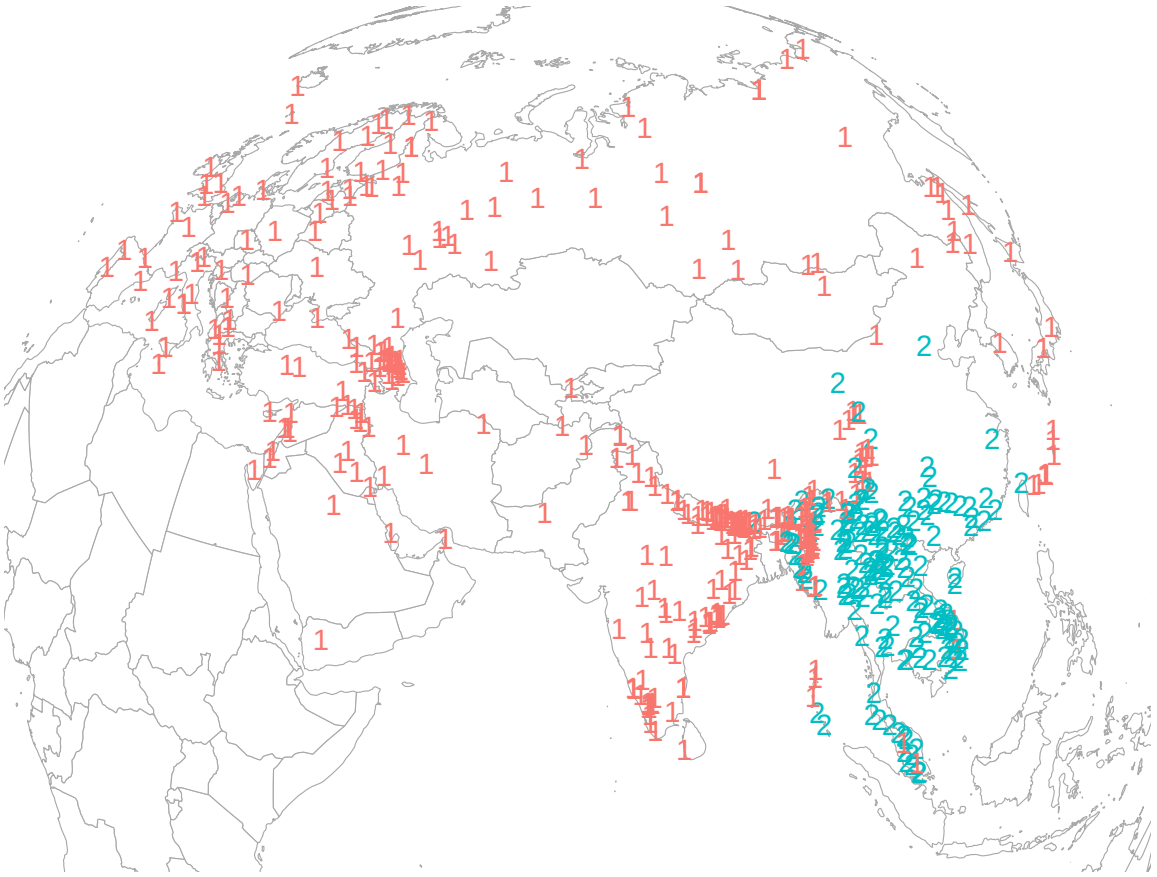
K2 map.

```
GramK2 <- EurasiaMap +
  geom_text(aes(x = Longitude,
                y = Latitude,
                label = K2,
                color = K2),
            data = GramK,
           show.legend = FALSE) +
  theme(legend.position = 'bottom')

cairo_pdf(file = "GramK2.pdf",
          family = "Times New Roman",
          width = 7,
          height = 5)
  GramK2
dev.off()
```

```
## pdf
##   2
```

```
GramK2
```



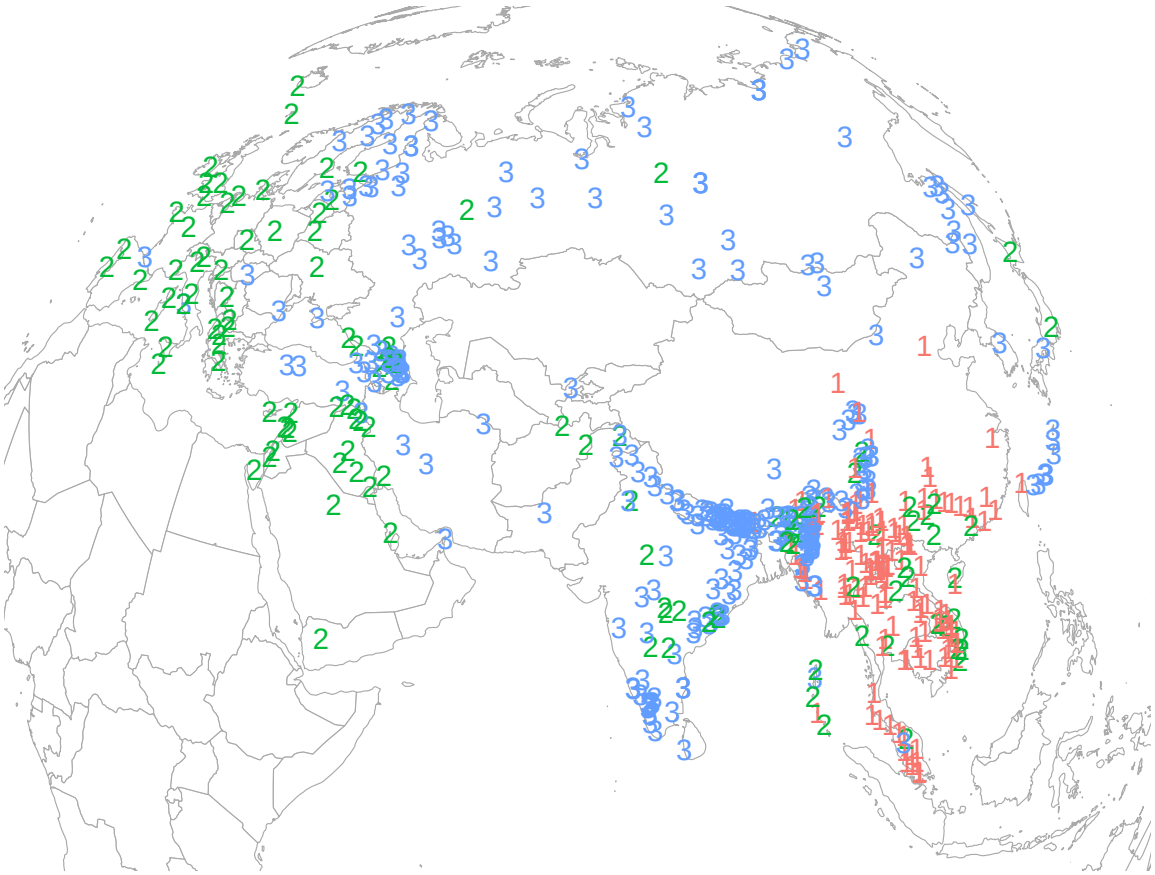
K3 map.

```
GramK3 <- EurasiaMap +
  geom_text(aes(x = Longitude,
                y = Latitude,
                label = K3,
                color = K3),
            data = GramK,
           show.legend = FALSE) +
  theme(legend.position = 'bottom')

cairo_pdf(file = "GramK3.pdf",
          family = "Times New Roman",
          width = 7,
          height = 5)
  GramK3
dev.off()
```

```
## pdf
##   2
```

```
GramK3
```



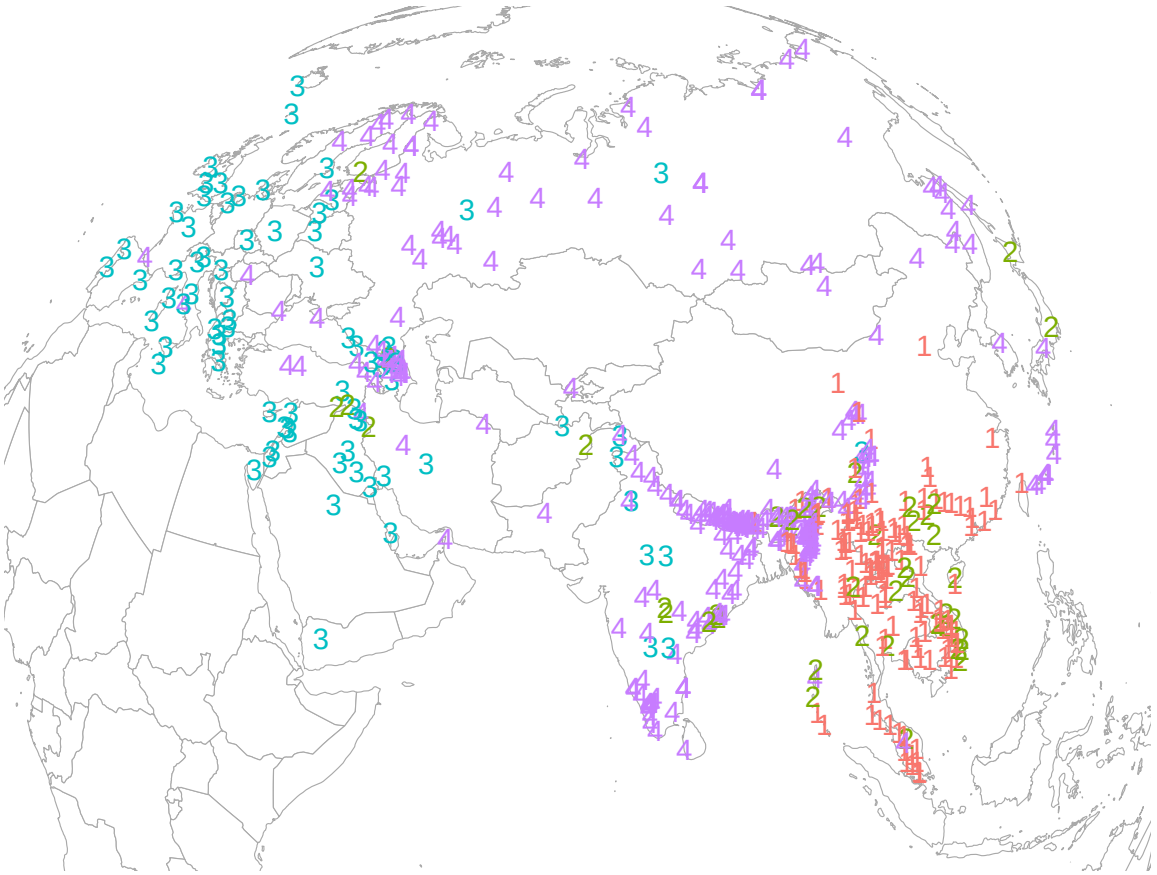
K4 map.

```
GramK4 <- EurasiaMap +
  geom_text(aes(x = Longitude,
                y = Latitude,
                label = K4,
                color = K4),
            data = GramK,
           show.legend = FALSE) +
  theme(legend.position = 'bottom')

cairo_pdf(file = "GramK4.pdf",
          family = "Times New Roman",
          width = 7,
          height = 5)
  GramK4
dev.off()
```

```
## pdf
##   2
```

```
GramK4
```



The Glottocodes of the intersecting lects.

```
Intersect <- Phonotacticon %>%
  .[Lect %in% Eurasia$Lect] %>%
  .[Glottocode %in% GrambankBinary$Language_ID] %>%
  .[, .(Glottocode, Lect)]

Intersect
```

```
##     Glottocode          Lect
##   1:   aoua1234          A'ou
##   2:   akaj1239       Akajeru
##   3:   amdo1237   Amdo Tibetan
##   4:   anga1288   Angami Naga
##   5:   aona1235       Ao Naga
##  ---
## 218:   yong1276 Yongbei Zhuang
## 219:   youl1235    Youle Jinuo
## 220:   yuec1235   Yue Chinese
## 221:   zaiw1241         Zaiwa
## 222:   zeme1240     Zeme Naga
```

Subset PhonoDist into intersecting lects and widen it.

```
PhonoDistIntersect <- PhonoDist %>%
  .[Lect_vs_Lect, on = .(Lect_vs_Lect)] %>%
  .[Lect %in% Intersect$Lect & i.Lect %in% Intersect$Lect] %>%
  .[, .(Lect, i.Lect, Distance)] %>%
  dcast(Lect ~ i.Lect, value.var = 'Distance') %>%
  .[, !c('Lect')] %>%
  as.matrix() %>%
  unname() %>%
  as.dist()

PhonoDistIntersect[1:10]
```

```
##  [1] 2.54 5.28 2.80 2.43 5.70 3.77 5.02 5.73 5.07 5.81
```

Subset GramDist into intersecting lects and widen it.

```
GramDistIntersect <- GrambankBinary %>%
  .[Intersect, on = c(Language_ID = 'Glottocode'), nomatch = NULL] %>%
  .[, Language_ID := NULL] %>%
  dcast(Lect ~ Parameter_ID, value.var = 'Value', fill = 0) %>%
  .[, Lect := NULL] %>%
  dist(method = 'manhattan')

GramDistIntersect[1:10]
```

```
##  [1] 158 116 127 119 154 152 111  58 137 155
```

Mantel test

```
mantel(PhonoDistIntersect, GramDistIntersect, method = 'pearson')
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = PhonoDistIntersect, ydis = GramDistIntersect, method = "pearson")
##
## Mantel statistic r: 0.215
##       Significance: 0.001
##
## Upper quantiles of permutations (null model):
##    90%    95%  97.5%    99%
## 0.0397 0.0512 0.0598 0.0708
## Permutation: free
## Number of permutations: 999
```

Comparing phonological distances to the number of shared genealogical layers

Load Glottolog data.

```
Glottolog <- fread('glottolog-cldf-4.4/cldf/values.csv') %>%
 as.data.table()

Glottolog
```

```
##                          ID Language_ID      Parameter_ID
##      1:            more1255-level    more1255            level
##      2:         more1255-category    more1255         category
##      3:    more1255-classification    more1255    classification
##      4: more1255-subclassification    more1255 subclassification
##      5:              more1255-aes    more1255              aes
##     ---
## 155396:         agob1245-category    agob1245         category
## 155397:    agob1245-classification    agob1245    classification
## 155398: agob1245-subclassification    agob1245 subclassification
## 155399:              agob1245-aes    agob1245              aes
## 155400:              agob1245-med    agob1245              med
## 5 variables not shown: [Value, Code_ID, Comment, Source, codeReference]
```

Subset dialects.

```
Dialects <- Glottolog %>%
  .[Parameter_ID == 'category'] %>%
  .[Value == 'Dialect'] %>%
  .[, Language_ID]

head(Dialects)
```

```
## [1] "ngkr1236" "ngka1237" "ngkr1237" "bark1251" "ngkr1235" "ngkr1238"
```

Subset languages and their classification.

```
Classification <- Glottolog %>%
  .[!Language_ID %in% Dialects] %>%
  .[Parameter_ID == 'classification'] %>%
  .[, .(Language_ID, Value)]

Classification
```

```
##       Language_ID           Value
##    1:    more1255            <NA>
##    2:    mong1349             <NA>
##    3:    kolp1236            <NA>
##    4:    naml1239             <NA>
##    5:    tana1288            <NA>
##   ---
## 13100:    alac1239         kawe1237
## 13101:    qawa1238 kawe1237/nort1506
## 13102:    alac1240 kawe1237/nort1506
## 13103:    idii1243          paho1240
## 13104:    agob1244           paho1240
```

Split the classification column.

```
Classification_split <-
  str_split_fixed(Classification$Value, pattern = '/', n = Inf) %>%
  as.data.table()

colnames(Classification_split) <-
  c('Family', 1:(ncol(Classification_split) - 1)) %>% as.character()

Classification_split
```

```
##          Family       1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
##    1:     <NA>
```

```
##     2:     <NA>
##     3:     <NA>
##     4:     <NA>
##     5:     <NA>
##    ---
## 13100: kawe1237
## 13101: kawe1237 nort1506
## 13102: kawe1237 nort1506
## 13103: paho1240
## 13104: paho1240
##         23
##     1:
##     2:
##     3:
##     4:
##     5:
##    ---
## 13100:
## 13101:
## 13102:
## 13103:
## 13104:
## 2 variables not shown: [24, 25]
```

Subset languages from Glottolog.

```
GlottoLects <- Glottolog %>%
  .[Value == 'language'] %>%
  .[, Language_ID]

head(GlottoLects)
```

```
## [1] "kolp1236" "tana1288" "touo1238" "bert1248" "sius1254" "cent2045"
```

Melt the classification table and exclude bookkept and unclassified families.

```
GeneaLects <-
  Classification_split %>%
  .[, Lect := Classification$Language_ID] %>%
  .[Lect %in% GlottoLects] %>%
  melt(id.vars = c('Lect', 'Family'),
       variable.name = 'Order',
       value.name = 'Branch') %>%
  .[!(Order != 1 & Branch == '')] %>%
  .[!Family %in% c('book1242', 'uncl1493')] %>%
  .[Family != '<NA>'] %>%
  .[Branch == '', Branch := Lect]

GeneaLects
```

Calculate the number of shared layers.

```
SharedLayers <- GeneaLects %>%
  .[GeneaLects, on = .(Family, Order), allow.cartesian = TRUE] %>%
  .[Lect != i.Lect] %>%
  .[, Shared := Branch == i.Branch] %>%
  .[, .(Lect, i.Lect, Shared)] %>%
  .[, .(SharedLayers = sum(Shared)), by = .(Lect, i.Lect)]

SharedLayers
```

```
##             Lect   i.Lect SharedLayers
##       1: sota1242 smar1235            1
##       2: ngka1235 smar1235            1
##       3: badi1247 smar1235            1
##       4: yeii1239 smar1235            0
##       5: namb1293 smar1235            0
##      ---
## 4668118: alac1240 qawa1238            1
## 4668119: alac1239 alac1240            0
## 4668120: qawa1238 alac1240            1
## 4668121: agob1244 idii1243            0
## 4668122: idii1243 agob1244            0
```

Make a list of Glottocodes of Phonotacticon.

```
PhonoGlottocodes <- Phonotacticon %>%
  .[, .(Lect, Glottocode, Family)]

PhonoGlottocodes
```

```
##             Lect Glottocode           Family
##   1:       A'ou   aoua1234        Tai-Kadai
##   2:      Abaza   abaz1241     Abkhaz-Adyge
##   3:     Abkhaz   abkh1244     Abkhaz-Adyge
##   4:     Adyghe   adyg1241     Abkhaz-Adyge
##   5:    Akajeru   akaj1239 Great Andamanese
##  ---
## 512: Yue Chinese   yuec1235     Sino-Tibetan
## 513:      Zaiwa   zaiw1241     Sino-Tibetan
## 514:     Zauzou   zauz1238     Sino-Tibetan
## 515:        Zbu   zbua1234     Sino-Tibetan
## 516:   Zeme Naga   zeme1240     Sino-Tibetan
```

Make a table of phonological distances and the number of shared layers.

```
PhonoGenea <-
  Lect_iLect %>%
  .[PhonoGlottocodes, on = .(Lect), nomatch = 0] %>%
  .[PhonoGlottocodes, on = .(i.Lect = Lect), nomatch = 0] %>%
  .[, Distance := PhonoDist$Distance] %>%
  .[Family == i.Family & Lect != i.Lect] %>%
  .[SharedLayers, on = .(Glottocode = Lect, i.Glottocode = i.Lect), nomatch = 0] %>%
  .[, .(Lect, i.Lect, Family, Distance, SharedLayers)]

PhonoGenea
```

```
##                Lect             i.Lect             Family Distance
##     1:          Bonan              Dagur     Mongolic-Khitan     1.96
##     2:          Dagur       Halh Mongolian     Mongolic-Khitan     2.31
##     3:       Dongxiang       Halh Mongolian     Mongolic-Khitan     2.65
##     4:          Bonan       Halh Mongolian     Mongolic-Khitan     3.79
##     5:          Dagur Peripheral Mongolian     Mongolic-Khitan     2.84
##    ---
## 10431:     Forest Enets          Pite Saami             Uralic     1.63
## 10432:          Ingrian          Pite Saami             Uralic     2.88
## 10433: Northern Yukaghir    Southern Yukaghir           Yukaghir     5.18
## 10434:    Cypriot Arabic         Neo-Mandaic       Afro-Asiatic     3.12
## 10435:          Chukchi             Koryak Chukotko-Kamchatkan     1.51
## 1 variable not shown: [SharedLayers]
```

Subset families that have at least one layer and at least three sample lect pairs.

```
PhonoGeneaFamily <-
  PhonoGenea %>%
  .[, Max := max(SharedLayers), by = .(Family)] %>%
  .[Max >= 1] %>%
  .[, .N, by = .(Family)] %>%
  .[N >= 3] %>%
  .[, .(Family)] %>%
  pull(Family)

PhonoGeneaFamily
```

```
##  [1] "Mongolic-Khitan"   "Tai-Kadai"         "Austronesian"
##  [4] "Tungusic"          "Hmong-Mien"        "Nakh-Daghestanian"
##  [7] "Turkic"            "Japonic"           "Indo-European"
## [10] "Dravidian"         "Austroasiatic"     "Sino-Tibetan"
## [13] "Uralic"
```

Test the correlation between the phonological distance and the number of shared layers.

```
PhonoGeneaPearson <- PhonoGenea %>%
 .[Family %in% PhonoGeneaFamily] %>%
 .[, cor.test(Distance, SharedLayers), by = Family] %>%
 .[, .(Family, parameter, estimate, p.value)] %>%
 .[!duplicated(.)] %>%
 .[, parameter := as.integer(parameter + 2)] %>%
 .[, p.value := p.adjust(p.value)] %>%
 .[order(estimate)]

colnames(PhonoGeneaPearson) <-
 c('Family', 'Number of lect pairs', 'r', 'FDR')

PhonoGeneaPearson %>%
 xtable(type = 'latex',
        label = 'PhonoGenea',
        caption = 'Pearson\'s correlation efficient (r) and the false discovery rate (FDR) between the phonological dista
 booktabs('PhonoGenea.tex')

PhonoGeneaPearson
```

```
##              Family Number of lect pairs       r FDR
##  1: Nakh-Daghestanian                 78 -0.1944   1
##  2:           Turkic                 66 -0.1801   1
##  3:        Tai-Kadai                153 -0.0933   1
##  4:        Dravidian                210 -0.0684   1
##  5:    Indo-European                1953 -0.0284   1
## ---
##  9:           Uralic                 36  0.1173   1
## 10:          Japonic                  6  0.1321   1
## 11:         Tungusic                 15  0.3380   1
## 12:      Austronesian                 21  0.3870   1
## 13:       Hmong-Mien                  10  0.4216   1
```

Convert the matrix of morphosyntactic distance into a data.table.

```
GramDistDT <- GramDist %>%
 as.matrix() %>%
 as.data.table()

colnames(GramDistDT) <- GrambankLects$ID

GramDistDT
```

```
##     abkh1244 acha1249 aghw1237 aheu1239 ahom1240 ainu1240 aito1238 akab1249
##  1:        0      153      149      148      156      122      180      122
##  2:      153        0      162       53       73      127      131      125
##  3:      149      162        0      141      141      163      149      137
```

```
##   4:     148      53     141       0      58     124     118     136
##   5:     156      73     141      58       0     150     118     132
##  ---
## 559:     134      77     165      86      96     122     138     130
## 560:     139     128     138     133     131     131     149     121
## 561:     145      60     134      51      69     133     127     123
## 562:     139      88     132     103      99     145     139     111
## 563:     144      99     131     100     106     124     122     128
## 555 variables not shown: [akaj1239, akha1245, akka1240, aluk1238, amdo1237, amri1238, anci1242, anci1244, ar
```

Retrieve the corresponding names of Glottocodes.

```
GlottologNames <- fread('glottolog-cldf-4.4/cldf/languages.csv') %>%
 as.data.table() %>%
 .[, .(ID, Name)]

GlottologNames
```

```
##             ID                  Name
##     1: more1255                   Yam
##     2: mong1349        Mongolic-Khitan
##     3: kolp1236  Kol (Papua New Guinea)
##     4: naml1239          Namla-Tofanma
##     5: tana1288            Tanahmerah
##    ---
## 25896: tame1238                 Taeme
## 25897: nucl1597                  Idi
## 25898: kawa1281                Kawam
## 25899: ende1235 Ende (Papua New Guinea)
## 25900: agob1245                 Agöb
```

Make a table of the Eurasian lects of Grambank and their family.

```
GramFamily <- Glottolog %>%
 .[Language_ID %in% Grambank$Language_ID &
    Parameter_ID == 'classification'] %>%
 .[, .(Language_ID, Value)] %>%
 .[Value == '<NA>', Value := Language_ID] %>%
 .[, Value := sapply(strsplit(Value, "/"), function(x) x[1])] %>%
 .[GlottologNames, on = .(Value = ID), nomatch = 0] %>%
 .[, .(Language_ID, Name)]

colnames(GramFamily) <- c('Lect', 'Family')

GramFamily
```

```
##         Lect           Family
```

```
##   1: daur1238     Mongolic-Khitan
##   2: halh1238     Mongolic-Khitan
##   3: kalm1243     Mongolic-Khitan
##   4: peri1253     Mongolic-Khitan
##   5: russ1264     Mongolic-Khitan
##  ---
## 559: hert1241        Afro-Asiatic
## 560: mlah1239        Afro-Asiatic
## 561: nucl1706        Afro-Asiatic
## 562: west2763        Afro-Asiatic
## 563: chuk1273 Chukotko-Kamchatkan
```

Make a table of the morphosyntactic distance and the number of shared layers (while also excluding the sign language "family' ', because it's not really a family but rather a modality)

```
GramGenea <- GramDistDT %>%
 .[, Lect := GrambankLects$ID] %>%
 melt(id.vars = 'Lect',
      variable.name = 'i.Lect',
      value.name = 'Distance') %>%
 .[, Lect_vs_Lect := str_c(pmin(as.character(Lect), as.character(i.Lect)),
                           'vs.',
                           pmax(as.character(Lect), as.character(i.Lect)),
                          sep = ' ')] %>%
 .[!duplicated(Lect_vs_Lect)] %>%
 .[GramFamily, on = .(Lect)] %>%
 .[GramFamily, on = .(i.Lect = Lect)] %>%
 .[Family != 'Sign Language' &
     Family == i.Family &
     Lect != i.Lect] %>%
 .[, .(Lect, i.Lect, Family, Distance)] %>%
 .[SharedLayers, on = .(Lect, i.Lect), nomatch = 0]

GramGenea
```

```
##           Lect   i.Lect         Family Distance SharedLayers
##     1: halh1238 daur1238 Mongolic-Khitan       77           1
##     2: kalm1243 daur1238 Mongolic-Khitan       82           1
##     3: peri1253 daur1238 Mongolic-Khitan       55           1
##     4: russ1264 daur1238 Mongolic-Khitan       73           1
##     5: dong1285 daur1238 Mongolic-Khitan       74           1
##    ---
## 23908: stan1318 nucl1706    Afro-Asiatic      104           3
## 23909: sana1295 nucl1706    Afro-Asiatic      116           3
## 23910: ugar1238 nucl1706    Afro-Asiatic      105           4
## 23911: phoe1239 nucl1706    Afro-Asiatic      110           4
## 23912: west2763 nucl1706    Afro-Asiatic      115           7
```

Subset families of GramGenea to those with at least one layer and at least three lect pairs.

```
GramGeneaFamilies <-
 GramGenea %>%
 .[, Max := max(SharedLayers), by = .(Family)] %>%
 .[Max >= 1] %>%
 .[, .N, by = .(Family)] %>%
 .[N >= 3] %>%
 .[, .(Family)] %>%
 pull(Family) %>%
 unique()

GramGeneaFamilies
```

```
##  [1] "Mongolic-Khitan"   "Tai-Kadai"         "Austronesian"
##  [4] "Tungusic"          "Hmong-Mien"        "Nakh-Daghestanian"
##  [7] "Turkic"            "Japonic"           "Indo-European"
## [10] "Dravidian"         "Austroasiatic"     "Sino-Tibetan"
## [13] "Uralic"            "Afro-Asiatic"
```

Test the correlation between the morphosyntactic distance and the number of shared layers.

```
GramGeneaPearson <- GramGenea %>%
 .[Family %in% GramGeneaFamilies] %>%
 .[, cor.test(Distance, SharedLayers), by = Family] %>%
 .[, .(Family, parameter, estimate, p.value)] %>%
 .[!duplicated(.)] %>%
 .[, parameter := as.integer(parameter + 2)] %>%
 .[, p.value := p.adjust(p.value)] %>%
 .[order(estimate)]

colnames(GramGeneaPearson) <-
 c('Family', 'Number of lect pairs', 'r', 'FDR')

GramGeneaPearson %>%
 xtable(type = 'latex',
        label = 'GramGenea',
        caption = 'Pearson\'s correlation efficient (r) and the false discovery rate (FDR) between the morphosyntactic
 booktabs('GramGenea.tex')

GramGeneaPearson
```

```
##              Family Number of lect pairs       r       FDR
##  1:         Japonic                  15 -0.7487  1.06e-02
##  2:          Uralic                 300 -0.5707  2.54e-26
##  3:        Dravidian                 435 -0.5097  4.61e-29
##  4:    Indo-European                2016 -0.4635 8.56e-107
```

```
##  5:           Turkic                91 -0.3701  2.75e-03
## ---
## 10:         Tai-Kadai               120 -0.1408  6.25e-01
## 11: Nakh-Daghestanian               210 -0.1299  4.22e-01
## 12:       Afro-Asiatic              231 -0.0749  7.78e-01
## 13:        Hmong-Mien                45  0.0489  7.78e-01
## 14:       Austronesian              120  0.1193  7.78e-01
```