University of Zurich

Master's Thesis - GEO 511

Global Analysis of the Influence of Geographical Factors on Contact-Induced Language Change

---

Fabiola Kälin
(11-733-730)

Supervisors

Dr. Curdin Derungs

Prof. Dr. Balthasar Bickel
Department of Comparative Linguistics
Plattenstrasse 54
8032 Zürich
balthasar.bickel@uzh.ch

Prof. Dr. Robert Weibel
(Faculty Member)

Department of Geography
Geographic Information Systems

Submission: January 27, 2017

Abstract

Languages are in constant change, whereby a general distinction is made between contact-induced and internally-driven change. Intense contact between languages is associated with a fast and profound language change. As contact situations emerge in geographic space, geography influences the probability of contact. Based on theories originating from research on language diversity, this thesis performs a global analysis of contact-induced language change from a geographical perspective. The analysis is carried out on the scale of language families, i.e. language change is assessed for language families and geographical factors are computed for their respective areas.

The contribution of this thesis is twofold: firstly, it shows how geographical factors that influence contact-induced language change can be modelled. Secondly, it provides new insights into the influence of geography on contact-induced language change. The results suggest that climatic and topographical characteristics, the number of neighbours and shape compactness of a language family area influence language change. These factors favour or disfavour the emergence of language contact and they influence the effectiveness of contact in contact situations. The results further suggest that the sound system (phonology) of a language changes more rapidly in a contact situation than its grammar.

Acknowledgements

Many thanks to my supervisors Dr. Curdin Derungs, Prof. Dr. Balthasar Bickel and Prof. Dr. Robert Weibel for their continuous support, the helpful meetings, and especially for their flexibility and great inputs.

Special thanks to Annabelle Jaggi, Dina Tageldin, Anne Wegmann, Larissa Kessler, Dorothee Kälin, Daniel Zimmermann and Ariana Dragusha for their inputs, help in Latex and proof-reading!

Further, I would like to thank Corin Meier for the good company at G10, at both weekends and late nights. As we all know: a sorrow shared is a sorrow halved.

Last but not least, special thanks to my family for supporting me throughout my whole life and especially throughout my years of study!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the present-day world people speak roughly 7'000 different languages, reflecting a cultural evolution of thousands of years (Greenhill et al., 2010). As languages are in constant change, modern languages provide insights into the past and allow for a partial reconstruction of the (pre)historic age. To do that, the distribution of languages, loanwords and similarities of features among languages are analysed. To date, however, a general theory of language change does not exist because it is still uncertain how and why languages change (Bowern and Evans, 2015). Although actual reasons for language change are difficult to assess, a general distinction is made between externally-motivated, i.e. contact-induced, and internally-driven language change, e.g. innovations within a speech community (Bowern, 2013). Intense language contact between different speech communities is associated with a fast and profound change of a language, whereas changes within isolated languages tend to occur rather slowly (Lucas, 2014). Furthermore, it is claimed that changes that are already in progress in a language can be accelerated when a contact situation emerges (Trudgill, 2011). In this thesis, the assumed association of language contact with rapid and profound language change is referred to as the contact hypothesis. The contact hypothesis forms the theoretical basis of the thesis.

As contact situations emerge in geographic space, geography influences the probability of contact. To date, several theories exist about the influence of geography on the mobility and contact of groups of people. Besides political and economic factors, topography and climate are seen as guides of movement and reasons for contact between groups. Large mountain ranges restrict the movement of groups and narrow the chance of contact with other populations (Stepp, Castaneda, and Cervone, 2005). Climate, on the one hand, influences the direction of long-term mobility as groups tend to move in the direction of stable ecological circumstances (Diamond, 1997). On the other hand, climatic conditions determine to what extent a group depends on establishing networks to other groups, which is referred to as ecological risk (Nettle, 1998).

Contact-induced language change and geography have only rarely been linked. In contact linguistics there has been a lot of research focusing on the linguistic mechanisms of contact-induced language change. To scholars it has been of great interest which historical, sociolinguistic, socioeconomic and political situations lead to language change (Thomason,

2001). The contact of groups has mainly been studied with the aim of explaining language diversity (for an overview see Gavin et al. (2013)), but it has so far not been tested for contact-induced language change.

In collaboration with the Department of Comparative Linguistics of the University of Zurich, this thesis systematically investigates contact-induced language change using methods from Geographic Information Science. The first aim is to gain new insights into the influence of geography on contact-induced language change on a global scale. In doing so, geography is seen as a space in which contact situations emerge. The characteristics of space may favour or disfavour the contact of groups. The second aim of this thesis is to find appropriate geographical factors for measuring the influence of geography on contact-induced language change.

1.1 Research Questions

The first aim of this thesis is reflected in the principle research question:

RQ 1 Do geographical factors influence language change?

This question investigates which geographical factors influence language change. To tackle this, language change is assessed on the level of language families. The geographical settings of language families are modelled along different factors such as topographical layout and climatic conditions, neighbourhood measures quantifying contact potential and geometric properties describing the spread of language families. Language change is assessed by analysing differences between languages of the same language families. This is done by estimating the rate of change of linguistic features of the structural domains of a language, namely phonology (sound system) and grammar (word and sentence structure).

The following sub-questions guide the analysis:

RQ 1.1 Does geography influence grammatical and phonological language change in a similar way?

This question investigates by which geographical factors the two domains are influenced and if one of the domains is influenced more strongly by language contact than the other. This issue is addressed by comparing the average rate of change of grammar and phonology per family with the different geographical factors.

RQ 1.2 The change of which linguistic features can be explained by geography?

This question addresses a different level of analysis, namely the feature level. It is investigated which linguistic features are influenced by which geographical factors. This issue is addressed by a pair-wise comparison of the rate of change of linguistic features with geographical factors.

RQ1 and its sub-questions are content-related. The second research question reflects the second aim of the thesis by addressing the methodology applied in this thesis:

RQ 2 How can geographical factors be determined for the investigation of language change?

This question addresses the suitability of the applied methodology for measuring contact-induced language change. It further investigates if the computed geographical factors are appropriate for measuring contact-induced language change on the granularity of language families.

1.2 Structure of the Thesis

The thesis is structured as follows: chapter 2 introduces basic linguistic concepts and contact-induced language change from a linguistic perspective. Subsequently, chapter 3 outlines theories and studies that have been carried out in the interdisciplinary field of linguistics and geography with a focus on language contact and geography. Chapter 4 describes the research gap, provides an overview of the approach applied to investigate contact-induced language change and hypotheses are proposed. Subsequently, chapter 5 describes the different datasets used for the analysis, i.e. the transition rate data and the linguistic database that is used for modelling language and language family areas, which is described in chapter 6. In this chapter the computation of the geographical factors is depicted as well. Chapter 7 sets out the results of the correlation analyses which are then interpreted chapter 8. Chapter 9 discusses the outcomes by answering the posed research questions and points out the major limitations of the applied approach. The conclusion is drawn in chapter 10 and an outlook is given in chapter 11.

Chapter 2

Languages and Language Change

This section introduces the linguistic concepts language and language family and describes their representation. The structural domains of language are introduced and basic theories of language contact are described. As the focus of this thesis lies on the spatial modelling, this section only serves as an introduction into the linguistic context of the thesis, in-depth discussions of concepts and theories are not provided.

2.1 Language and Language Family

The concepts language and language family are outlined in this section and distributional patterns across the globe are described. Subsequently, the representation of language families as trees is delineated.

2.1.1 Language

In linguistics, the most widely used indicator of what a language is and what dialects of a single language are, is the criterion of mutual intelligibility. This means that if two speakers are able to understand each other without having learnt the language of the other speaker, they speak dialects of the same language. If they do not understand each other, they speak different languages (McGregor, 2015), however this boundary is fuzzy. The transition of two dialects of a single language into two different languages, is greatly dependent on conversational contexts, social factors, etc., which decide whether or not speakers are able to understand each other (Thomason, 2001).

There are around 7'000 different languages spoken in the world, of which 400 are nearly extinct (McGregor, 2015). These languages are distributed unequally across the globe, following a latitudinal gradient. This means that language diversity is high in equatorial regions and relatively low in more northerly and southerly regions. The highest language density exists in New Guinea and its nearby islands, where more than 1000 languages are spoken (McGregor, 2015). Languages are very uneven in terms of their speaker population. The languages with the most native speakers are Mandarin Chinese, Japanese, Spanish, English, Hindi, Arabic, Portuguese, Bengali, Russian and Japanese. They are spoken by more

than 100 Million people each, making up over 40% of the world population together. At the other extreme, approximately 3'500 languages are spoken by less than 10'000 speakers each, making up less than 0.3% of the world population together (McGregor, 2015).

2.1.2 Language Family

Languages that derive from a single ancestor, a so-called proto-language, are genetically related. This proto-language has split into a range of varieties that have become mutually unintelligible over time. This group of languages is considered to be a single language family (McGregor, 2015). Languages of the same family evince sound correspondences and share structural features and cognates (words with the same origin) (Nichols, 1990). The maximum time depth of assessing genetic relationships is about 10'000 years. Tracing back genetic relationship to a more distant past is not possible because change is too rapid (McGregor, 2015). Special cases of families are languages with no known genetic relatives, such as the language Burushaski which is spoken in Pakistan. Such languages are called isolates (McGregor, 2015).

The distribution of language families across the globe reflects the distributions of the languages; in the equatorial regions, the density of different language families is higher than in more northern or southern regions (Nichols, 1990). For instance, the Glottolog database (Hammarström et al., 2016) identifies more than 120 language families in New Guinea and its nearby islands. A further important fact is that language families are very unequal in size and speaker population; the biggest language families are Indo-European, Austronesian, Afro-Asiatic, Niger-Congo, Sino-Tibetan and Trans-New-Guinea. These families contain more than 300 languages each, making up 60% of all languages accounting for 80% of the world population. On the other extreme, there are several isolates and a lot of rather small families containing only a few languages with a few thousand speakers.

Language families are often represented as trees, which is described in the following section.

2.1.3 Language Family Tree

Language family trees show the evolutionary relations between the members of a language family (McGregor, 2015). The tree representation is a useful descriptive method, but for a lot of families, detailed lineages are disputed and some languages cannot be positioned within the family tree (Dixon, 1997). Moreover, the tree representation is not suitable for the representation of different types of language change, e.g. language contact (Enfield, 2005). Figure 2.1 depicts a simplified version of the language family tree of Indo-European with "Proto-Indo-European", the ancestor, on the top. The languages marked in grey are ancestors of the modern languages marked in black.

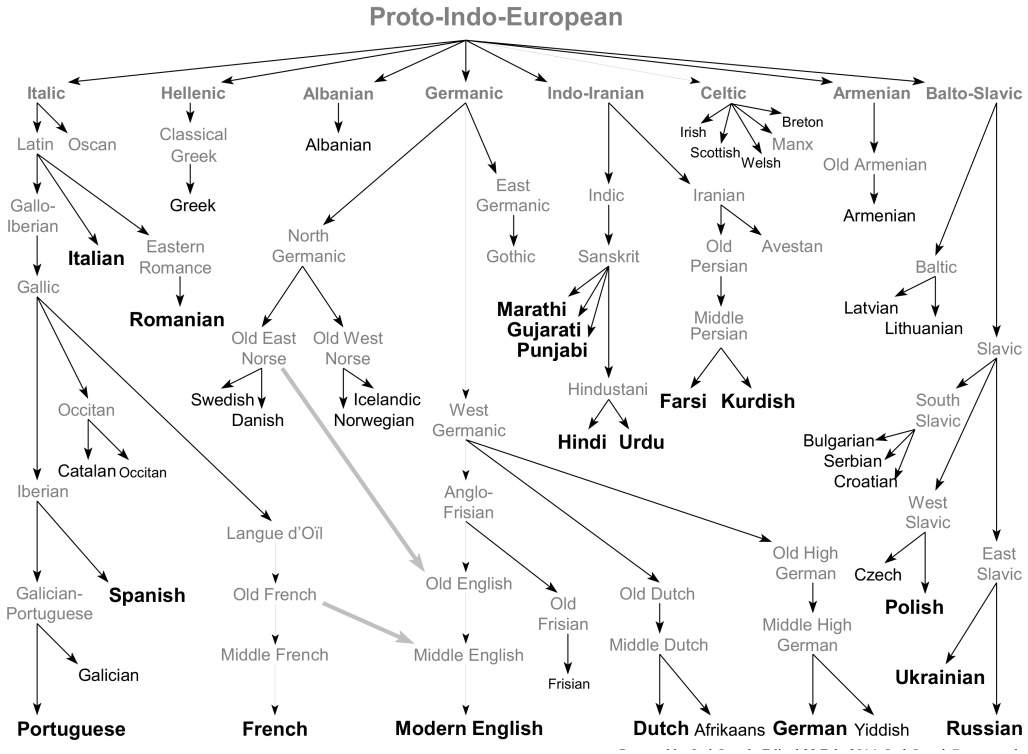


Figure 2.1: Simplified language family tree of Indo-European. Ancient languages are marked in grey, modern languages in black. The grey arrows do not indicate direct descent, but the influence of a language on another. *Figure prepared by Jack Lynch (available at: http://andromeda.rutgers.edu/~jlynch/language.html, accessed on 18.01.2017)*

In literature, language families and stocks are usually differentiated as two levels of genetic groupings. A stock is a hypothetical super group of language families and thus has a bigger time depth and the languages show fewer similarities (Nichols, 1990). However, the number of families and stocks is disputed, because clear boundaries cannot be drawn between the two levels. Dependent on the definition, criteria and methods for establishing families, the number of groupings varies considerably. The Glottolog database, for instance, identifies roughly 420 families (Hammarström et al., 2016). The discrepancies are also reflected in the linguistic databases used in this thesis in the sense that the genetic groupings of languages do not match perfectly. In this thesis, the top-level groupings of languages are referred to as language families although it might be argued that some groupings are a stock rather than a family.

Language family trees are established in historical linguistics. A modern method to estimate language family trees are phylogenetic methods originating from biology. The application of such computational techniques on linguistic data has increased over the last two centuries due to many parallels between biological evolution and the evolution of languages (Atkinson and Gray, 2005). These data-driven quantitative methods aim at estimating phylogenies, i.e. evolutionary histories of language families (Nichols and Warnow, 2008). Most of them model the historical behaviour of words or morphemes (see explanation in the next section) with the same origin (cognates). The likeliest tree is then selected as the tree of a

language family (Dunn, 2015). In this thesis, phylogenetic trees are used for inferring the transition rates of linguistic features. This will be outlined in more detail in chapter 5. The two domains of which transition rates are estimated are introduced in the following section.

2.2 Language Structure

This section provides a short overview of the domains making up the structure of a language, namely phonology and grammar. Further important aspects that characterise a language are semantics and pragmatics dealing with meaning and lexicon dealing with the classification of words into parts-of-speech. They are not introduced further because they are not incorporated in this thesis.

2.2.1 Phonology

Phonology deals with the sound system of languages; it is concerned with the systematic patterning of sounds in a language and analyses the characteristics that are significant in a sound system of a language. The most important concept in phonology is the phoneme. This is a distinctive sound of a language that is able to differentiate between words in a language, or to change the meaning of a word respectively. The inventory of phonemes in a language is also examined by phonology (McGregor, 2015).

2.2.2 Grammar

Grammar is constituted by morphology and syntax, i.e. the internal make-up of words and how these words are put together to form a sentence. Morphology is the study of words dealing with the structure and function of word forms (Zeige, 2015). The smallest meaningful unit of a language is the morpheme, which is the most important concept in morphology. Morphemes are combined in order to form a word, for example, *unlikely* is a sequence of three morphemes, namely *un-*, *like* and *-ly*. The ordering of morphemes usually follows regularities, which enable a general characterisation of the morphological form of words of particular types, such as nouns (McGregor, 2015).

Syntax deals with how words can be put together to form larger units, such as clauses and phrases, that can again be combined to build sentences. The sentence is crucial to syntax because it is the biggest unit in a language that is grammatically patterned. The structure of these units, e.g., the order of subject, verb and object, differs across languages (McGregor, 2015). For instance, word order in Latin was relatively free, i.e. subject, verb, and object could be put together in several ways. The Romance languages descending from Latin (see figure 2.1), however, have a fixed word order (subject-verb-object). These languages have a simplified morphology compared to Latin, as they do not show case marking of

nouns. In order to distinguish the subject from the object the word order has been rigidified (McGregor, 2015).

This section has introduced the two domains of language structure that are incorporated in this thesis. The next section provides the context in which grammar and phonology are examined, namely language contact and language change.

2.3 Language Contact and Change

Until now, a general theory of how and why languages change does not exist (Bowern and Evans, 2015), however many approaches have been put forward to explain language change. For example, on the one hand, languages may change in order to optimise the fit between the linguistic and the biological system, which processes the linguistic system. On the other hand, languages also change to maximise the fit between the linguistic system and the communication demands speakers have (Bickel, 2015). It has also been proposed that major changes coincide with periods of fundamental social change influencing communication networks. A further important cause of language change is the contact of groups speaking different languages (McGregor, 2015). Contact is often one of the causes leading to language change, however, change is often a result from multiple interacting causes, both external and internal ones (Thomason, 2001).

2.3.1 Languages in Contact

Thomason (2001) distinguishes between three incidents resulting from language contact: languages mix, become extinct or change due to a contact situation. These three incidents are addressed in the following. In some situations, so-called contact-languages emerge in case different speech communities do not learn each other's language which results in a mix of the languages in contact. Mixed languages strictly used as *lingua francas* (language of communication between speakers of different languages), are called pidgins (Kaye and Tosco, 2003). If such a mixed language becomes the main language of a community, it is called creole (Thomason, 2001). The second incident is the disappearance of a language. The most common reason for that is the shift of a speech community to another language. A further possibility is that it becomes extinct because all speakers die, for instance, when they are massacred by invaders. The third possibility is contact-induced language change, wherein the focus of this thesis lies. This is the most common result of language contact, whereby typically at least one language is influenced by at least one of the other languages (Thomason, 2001). In simple terms, language contact is "the use of more than one language in the same place at the same time" (Thomason, 2001, p. 1). As this is the case in a lot of places all over the world, language contact (and thus also changes provoked by language contact) is not exceptional, but rather normal. It is highly improbable that any language has developed in complete isolation from other languages. However, the intensity of the contact situation plays an important role, as it influences the changes occurring during

the contact situation (Thomason, 2001). Long-distance contacts, for example via religious languages, published texts, etc., are an important kind of language contact, but they are not of interest in this thesis. The changes occurring in contact-induced language change and the mechanisms that are provoked by contact situations are described in the following.

2.3.2 Contact-Induced Language Change

In general, any linguistic feature in a language can change over time, but not at constant rates (Atkinson et al., 2008). The existence of intrinsic stabilities of certain features across language families and geographical areas has been claimed, but there have been found counterexamples for every claim. Given the right combination of linguistic and social circumstances, any element of a language can be adopted by another language, but some items are more resistant to transfer than others (Thomason, 2001).

The most fundamental mechanisms provoked by language contact are interference on the one hand and convergence on the other. Interference is the transfer of structures and/or material from one language into another. This importation may occur by borrowing or by shift-induced interference (Thomason, 2001). Borrowing is the incorporation of a feature of a different language into a group's native language, whereby the native language is maintained (Thomason and Kaufman, 1988). The borrowing of words is the most common type of foreign influence. For instance, a lot of French words were integrated into English after the Normans conquered England in 1066 (Thomason, 2001) (see grey arrow in figure 2.1 indicating the influence of Old French on Middle English). Shift-induced interference is the result of imperfect learning of a group of speakers during the shift to another language (Thomason and Kaufman, 1988). Convergence, the second mechanism, is any process through which languages in contact situations become more similar. This mechanism can be described as mutual interference resulting in convergent structures of two or more languages with no single source. Convergence is the main mechanism in so-called linguistic areas (*Sprachbund* situations). A linguistic area is a geographical region within which languages share structural features as a result of language contact rather than as a result of genetic relation (Thomason, 2001).

Through such transfer processes, any linguistic feature can be adopted by a language. Regarding phonological and grammatical change, phonological features tend to be transferred more easily by contact-induced language change. The reason is that lexical borrowing (i.e. the borrowing of words), which is very likely to occur (Tadmor, 2009; Sankoff, 2002; Thomason, 2010), usually entails phonological changes due to subsequent adjustments in the phonology of the language adopting words. Such adjustments may not only be applied to borrowed vocabulary, but may also be applied to native lexicon (Sankoff, 2002).

2.3.3 Contact Hypothesis

As mentioned before, language contact is generally associated with a fast and profound change in at least one of the languages in contact (Lucas, 2014). On the one hand, contact situations can induce interference or convergence and lead to rapid changes in one or more of the languages in contact, depending on many other factors (addressed in the next section). On the other hand, a consequence of language contact is that it can speed up internally-motivated changes (e.g. due to inherently unstable aspects in a language) that are already in process (Trudgill, 2011). This does not necessarily result in interference or convergence, but may simply accelerate the ongoing change (Lucas, 2014). This rapid change evoked by contact situations contrasts the exceptionally slow language change of languages in almost total isolation (Lucas, 2014). Although most probably, no language has ever been totally isolated from other languages for more than one hundred years (Thomason, 2001), some languages developed in more isolation than others. The example of Icelandic suggests that isolation or lack of contact respectively promotes conservatism, resulting in slow language change (Lucas, 2014). An example of fast contact-induced language change is the integration of lexical items of French into the English language, as mentioned before. Not only words, but also morphemes such as *able* and *ment* were attached to native English words (McGregor, 2015).

This thesis is based on the contrasting rates of change of languages that are relatively isolated and of languages constantly in contact. As mentioned before, this is referred to as the contact hypothesis implying that rapid language change is induced by language contact. If and how fast languages in contact change, however, is dependent on many factors. Some of these are addressed in the next section.

2.3.4 Effectiveness of Language Contact

There are numerous political, socioeconomic and sociolinguistic factors that determine whether languages in contact situations change at all and if they do, how fast they change. One important factor is the intensity of a contact situation resulting in greater interference when contact is intense. Intensity is a matter of the amount of cultural pressure applied by one speech community on another. This is generally dependent on the duration of the contact, the size of the speaker populations and the socioeconomic dominance. The bigger the size of a group, the more probable is its socioeconomic dominance and the more probable that the smaller, subordinate group adopts features from the language spoken by the dominant group (Thomason, 2001).

The social identification of groups in contact with one another remains a crucial factor regarding the effectiveness of language contact. Languages are often regarded as an important dimension of social identity and may represent group categorization. They serve

as a promotion of identity, contrast, affiliation, power or solidarity (Sujoldžić and Muhvić-Dimanovski, 2004). This has an influence on the attitude of speakers which can either promote or hinder change. Thomason (2001) claims that speaker's attitudes are the main reason why contact-induced change remains unpredictable. These factors are crucial; however, they are not incorporated in this thesis, as the focus lies on geographical factors promoting or restricting the emergence of contact situations respectively.

This chapter provided an overview of basic linguistic concepts and has discussed language change from a linguistic perspective. The next chapter addresses the linkage of language contact and geography.

Chapter 3

Language Contact and Geography

This section starts by roughly introducing the interdisciplinary field of language contact and geography with a focus on research on mobility and the necessity of group contact.

Several studies have investigated the distribution of linguistic features in space. For instance, Nichols (1992) found distributions of structural features to be taking the form of global west-to-east clines and Atkinson (2011) observed that phonemic diversity declines with the increase of distance from the African continent. The processes underlying such spreads of linguistic features mainly include migration and interaction between groups. Mobility of groups thus plays an important role in language contact and besides being influenced by sociocultural factors, these processes are influenced by environmental (climatic and ecological) factors and topographical elements (Gavin et al., 2013). Theories about this subject matter mainly originate from research on species diversity in biology showing that spatial heterogeneity often correlates with greater diversity. For this, heterogeneity is quantified as habitat diversity or topographical complexity (see e.g. Kerr and Packer (1997)). Such theories have been adapted to explain linguistic diversity on the assumption that some processes leading to species diversity also lead to language diversity (Moore et al., 2002). The aim of these theories is to explain why groups separate or remain connected, whereby topographical elements are considered as restrictions on mobility (Stepp, Castaneda, and Cervone, 2005). Climatic conditions are seen as facilitating or impeding self-supply (Nettle, 1998) and as influencing livelihood strategies (Gavin et al., 2013). In these approaches geography defines circumstances that favour or disfavour contact between groups. These approaches can be adapted to investigate contact-induced language change because the process leading to high language diversity is the relative isolation of groups, which in contact linguistics is associated with low rates of language change. Another theory is about migration of people and claims that there is a latitudinal bias in large-scale migration patterns (Diamond, 1997). This theory is important for this thesis because it describes the preferred direction of long-term migration and it thus hypothesises how languages or language families, respectively, are spread and where language contact may happen. The following subsections introduce these three explanatory approaches in more detail and present some studies connected to them.

3.1 Topography and Mobility Restriction

The topography theory sees topographical features as potential restrictions on mobility. Mountain topography is thus a reason for the separation of communities, potentially leading to the isolation of groups of people because interaction with neighbouring groups is impeded (Stepp, Castaneda, and Cervone, 2005). In this regard, whether a landscape allows for transport or not may be a key factor (Nettle, 1996). Currie and Mace (2009) tested this theory for language diversity and Nichols (2014) conducted a study that is closely related to this topic.

Currie and Mace (2009) investigate several factors in order to explain language diversity. They found a significant correlation between the roughness of language areas (assessed by the standard deviation of the altitude) and the size of language areas. The correlation was positive, indicating that language area size is bigger in mountainous regions. This result opposes the topography theory, however, Currie and Mace (2009) do not take population size into account. In a lot of cases, mountainous areas are less populated than flat areas which might result in a low diversity although groups are rather isolated and do not speak the same languages.

Nichols (2014) shows the impact of geography on the development of languages of the Nakh-Daghestanian language family in the eastern half of the Great Caucasus range. Results show a positive correlation between linguistic structural properties of a language with the altitude at which a language is spoken. The grammar of languages spoken on high altitudes is more complex and certain sounds (so-called uvulars and ejective consonants) exist only in high altitude languages. Structural complexity in general also correlates with the sociolinguistic status of languages which is also dependent on altitude. Of course, altitude itself is not the immediate reason for structural complexity, but as peripheral locations like mountainous areas are precluded of large open networks, they are more easily isolated. Thus, mountain topography favours the isolation of ethnolinguistic groups and isolation, in turn, may favour linguistic complexity. Nichols'(2014) study does not show a direct connection to the contact hypothesis, but it suggests that topography is important for the isolation of languages and that isolated languages develop differently than languages in contact situations.

3.2 Climate and Subsistence

The subsistence strategy of a group is dependent on the ecological circumstances which in turn are influenced by climatic conditions. Diverse livelihood strategies may be developed and ethnolinguistic boundaries may be formed due to different habitats (Gavin et al., 2013).

Climate can also be interpreted as facilitating or impeding self-supply: Nettle (1998) explains the uneven distribution of languages across the globe (which is generally high in equatorial regions and lower in higher latitudes and arid regions) by the ecological risk. The

ecological risk describes how easy it is for a group to be self-subsistent, which depends on the amount of seasonal and/or inter-annual variation that is faced by people in their food supply (Nettle, 1996). This theory is closely related to other hypotheses predicting cultural diversity based on environmental factors such as productivity or biomass (Collard and Foley, 2002). The ecological risk can be described by climatic variability based on the mean temperature and precipitation throughout the year. In regions where climatic conditions allow for the production of food throughout the year, smaller groups of people tend to be autarkic because the ecological risk is low. This leads to a separation of groups into smaller social and economic communities because there is no need to establish large social networks across communities to deal with potential shortages of food. Without these networks, communities are less likely to be in contact with each other, although they are not necessarily isolated. This leads to less exchange between groups and language contact is less probable. As a consequence, in regions with a low ecological risk, language diversity is more pronounced due to the fragmentation into many small groups and languages. In regions with larger climate variability, and thus with an enhanced ecological risk, communities tend to establish larger social networks to secure their food supply, which results in more contact between groups and more widespread languages (Nettle, 1998). Although this theory was developed to describe language diversity, its argumentation is based on language contact, which makes it suitable for this thesis.

3.3 Climate and Large-Scale Migration

In Diamond's (1997) theory climate and ecology are accredited a fundamental role regarding the movement of human populations. The climate and ecological conditions respectively change less along latitudinal axes than along longitudinal axes. For instance, fauna and flora tend to remain the same. Thus, movement in east-west direction requires less adaption to new environments and there is no need to change the social system of a group. As a consequence, long-term historical migration in latitudinal direction is facilitated compared to the longitudinal direction (Diamond, 1997; Hammarström and Güldemann, 2014). Regarding linguistics, this theory implies that language contact and the spreading of linguistic features is more likely to occur in a latitudinal direction as well. A few studies linking linguistics to this theory have been published in recent years and are addressed in the following.

Güldemann (2010) explains the spread of large-scale linguistic areas of Africa on the one hand using Diamond's (1997) theory and on the other hand based on macro-topography. He distinguishes five linguistic areas, some of which show a strong latitudinal alignment while others rather extend in a north-south dimension. In accordance with Diamond's theory, the east-west oriented areas and their overall configuration relative to each other can be attributed to stable climatic characteristics. Furthermore, geophysical features such as mountains and water bodies also influence the large-scale distribution of typological features and impede a horizontal alignment of language areas. For instance, in eastern Africa,

linguistic features tend to be aligned in a north-south dimension. Güldemann (2010) attributes this to the East African Rift located at the eastern flank of Africa, which is the most notable topographical feature of the continent. The north-south alignment thus indicates that the movement of people was channelled along the Rift. Güldemann (2010) emphasizes that north-south migration also takes place, suggesting that trespassing the latitudinal boundary may lead to substantial change in typological characteristics of a language family. This may even result in a separate branch of this language family leading to higher structural diversity within a language family. Three aspects of Güldemann's (2010) findings are important for this thesis. First, linguistics features tend to cluster horizontally on a large scale due to the facilitated movement along latitudinal axes. Second, linguistic areas are also shaped by macro-topography. Although Güldemann (2010) does not speak of topography favouring isolation or contact, topography is important because it influences movement. The third aspect addresses accelerated change by entering a region with different ecological conditions.

Hammarström (2010) tested the Language Family Dispersal Hypothesis (LFDH) on a global scale. The LFDH is an approach to explain why some language families contain more languages than others. It is based on the hypothesis that the ancestral speakers of today's large language families had a technological advantage regarding agriculture, and therefore repressed groups speaking languages belonging to other families (Hammarström, 2010). For each family, ethnographic evidence was used to determine its subsistence status (agrarian vs. hunter-gatherer) and the geospatial spread of the family members was assessed and tested for horizontal alignment. Horizontality is defined as the ratio of the east-west to the north-south expansion of the coordinates of the family members. Based on the theory by Diamond (1997), families with the subsistence type farming were expected to be horizontally aligned. However, horizontality was not found to correlate for neither of the subsistence types nor for their combination but both the agrarian and hunter-gatherer families were found to tend to spread horizontally. This suggests that not only farming was responsible for the spread of some of the big families, but that other processes may have been involved as well. For this thesis, this means that language families tend to spread horizontally, which indicates that climatic factors influence the large-scale migration of groups of people.

Hammarström and Güldemann (2014) tested the alignment of the spread of word order and numeral systems on a global scale. To do that, they categorized languages into feature-wise homogeneous areas. It was tested if Köppen-Geiger climate zones and geophysical features influence the shape of these linguistic areas. In addition, their geospatial alignment was tested for horizontality. Hammerström and Güldemann's (2014) results show that Köppen-Geiger zones and geophysical features do not correlate with the shape of the linguistic macro-areas, although languages within Köppen-Geiger zones tend to be more homogeneous than random sets of a similar size. With respect to the horizontality of the areas, the results suggest that the bigger the size of the areal aggregations, the more horizontal their alignment. This coincides with the theory by Diamond (1997). It means that, on a small scale, people stay in similar climatic conditions in whichever direction they move,

but on a larger scale, the migration tends to occur along latitudinal axes (Hammarström and Güldemann, 2014). The finding important for this thesis is that, on a global scale, linguistic features are distributed horizontally due to the facilitated migration along latitudinal axes.

All of the reviewed studies indicate that climate and ecology are major factors influencing large-scale migration and thus also the large-scale distribution of linguistic features. The distribution of features is influenced, on the one hand, by the movement of groups per se, and, on the other hand, by the contact that is established with other groups as a result of this movement.

This chapter gave an overview of research in the interdisciplinary field of language contact and geography. The next chapter presents the research outline.

Chapter 4

Research Outline

The previous two chapters have introduced language contact from a linguistic and from a geographical perspective. This section outlines the research gap and the design of this thesis. Furthermore, it explains the expected impact of geographical factors on contact-induced language change.

4.1 Research Gap

On the one hand, there has been a lot of research focusing on linguistic mechanisms leading to contact-induced language change (see overview in section 2.3.2). Moreover, a lot of case-studies have been carried out in which sociolinguistic, socioeconomic, and political circumstances were observed to be fostering or impeding contact-induced language change (see overview in section 2.3.4). On the other hand, as described in chapter 3, several studies regarding geography and language diversity have been conducted, mainly addressing the topography theory and Nettle's (1998) ecological risk theory. The link between isolation or contact of groups and the speed of language change based on the linguistic contact hypothesis has been made, but it has not been tested to date. Diamond's (1997) theory has been linked to the horizontal alignment of linguistic features, but the impact of this theory on contact-induced language change has not been addressed. This thesis addresses this research gap and aims at contributing new insights into the influence of geography on contact-induced language change on a global level. The approach applied for the investigation of these theories is described in the following section.

4.2 Research Design

This thesis investigates the influence of language contact on language change using an exploratory approach. Languages are investigated in the context of their language family. This means that the investigation is carried out on the level of language families. Based on linguistic databases, language change is assessed on the one hand and language family areas are modelled on the other (see workflow in figure 4.1). The former is estimated by applying

phylogenetic methods to language family trees. In doing so, transition rates of linguistic features are computed for each language family. From the geographical perspective, languages are viewed as groups of people. A language family thus reflects several groups living in a geographical area with certain characteristics that favour or disfavour contact between groups. These characteristics are computed for the language family area using different geographical factors encompassing environmental characteristics, neighbourhood measures and geometric properties (explained in more detail in the next section). Subsequently, the influence of the geographical factors on language change is investigated. Language contact as a function of geography itself is not modelled but a preliminary analysis is performed in the form of a correlation analysis. In doing so, fast change is associated with language contact, based on the linguistic contact hypothesis.

Ideally, the development and divergence of languages would be compared to the geographical distribution of these languages over time, but this kind of spatiotemporal data is not available. Therefore, today's distribution of languages, which reflects their history of evolution, is taken as a basis for the computation of language family areas.

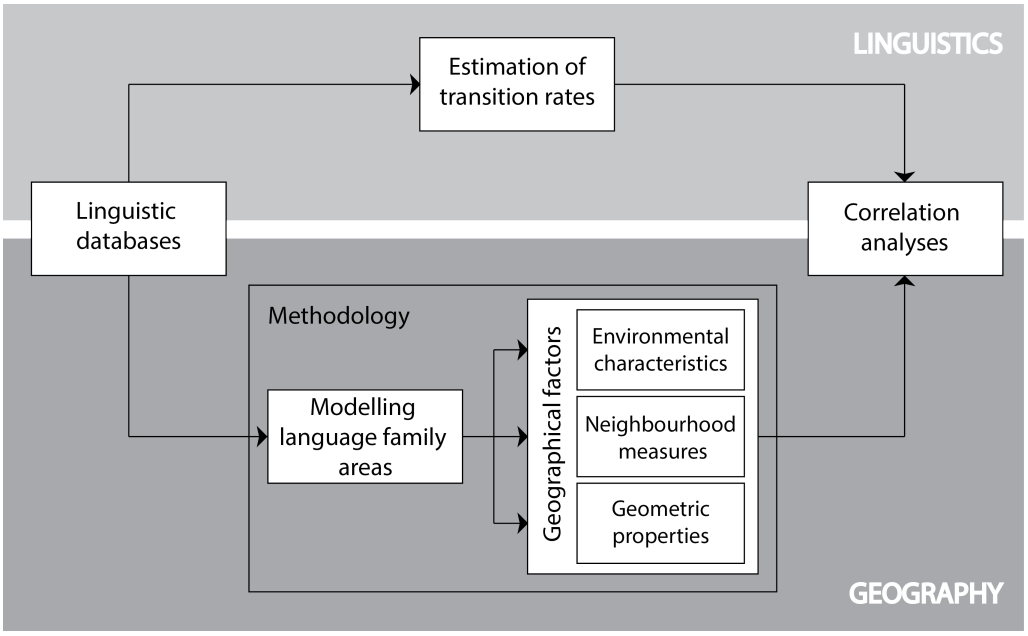


Figure 4.1: Approach applied in this thesis: based on the linguistic databases transition rates are estimated and language and language family areas are modelled. For these areas geographical factors are computed and subsequently compared to the transition rates in the form of correlation analyses.

4.3 Hypotheses

This section outlines the hypotheses connected to the research questions posed in section 1.1. As explained in chapter 2, phonological items of a language tend to be transferred more easily by contact-induced language change than grammatical items (Tadmor, 2009; Sankoff,

2002). Therefore, it is hypothesised that phonological change results in more and higher correlations with geographical factors than grammatical change.

Regarding contact-induced language change, there are two main types of conceptual factors influencing change: first, there are factors that drive the probability of contact, such as topographical characteristics and climatic conditions. Related to these factors is also contact potential, i.e. if there are ethnolinguistic groups with which contact is possible. This factor is treated as an own category. Second, there are factors that drive the probability of the effectiveness of contact. These factors involve sociocultural factors and geometry, whereby the focus lies on geometry in this thesis. In the following, the operationalisation of the conceptual factors is roughly described and it is addressed how the factors are expected to influence contact-induced language change.

4.3.1 Contact Probability

The probability of contact is influenced by environmental factors comprising climatic and topographical variables. Climate (temperature and precipitation) reflects the productivity of the environment and thus the necessity of contact between groups. High temperature and precipitation values lead to less contact among groups because self-sufficiency is facilitated (Nettle, 1998), which in turn is associated with a low rate of change due to minimal language contact. The variability of climatic conditions within a language family area, however, is associated with fundamental language change due to altered climatic conditions (Güldemann, 2010; Diamond, 1997). Migrating groups of people may have to restructure their social system to adapt to the new circumstances in order to survive. This in turn suggests a higher chance of contact to other groups leading to rapid language change. Topographical complexity reflects restriction of movement leading to isolation of groups (Stepp, Castaneda, and Cervone, 2005). This is associated with less language contact leading to a low rate of change.

4.3.2 Contact Potential

The number of neighbours maps the contact potential of a family, which is assessed on two levels, namely on the level of languages and on the level of language families (for details, see section 6.3.3). The expectation is that if there are a lot of neighbours, contact and thus contact-induced language change is more likely to happen. The correlation of geographic and phylogenetic proximity suggests that there is more language contact between genetically related languages than between unrelated languages (Bowern, 2013). Thus, in the case of language contact between groups of various families, differing linguistic features can be transferred. In the case of contact between languages of the same family the chance is higher that these features are similar already (Bowern, 2013). It is thus assumed, that contact with unrelated languages leads to more rapid language change.

4.3.3 Effectiveness of Contact

The geometry influencing the probability of effectiveness of language contact is operationalised in different ways. For instance, a large areal size of a language family suggests that intense contact between languages has occurred (Currie and Mace, 2009). Further, shape compactness and a horizontal spread of language family areas suggest stable social structures which in turn imply less fragmentation and less diversification (Diamond, 1997; Trudgill, 2010). This leads to more resistance against language change in contact situations.

In the following, the geographical factors quantifying the conceptual factors are referred to as environmental characteristics (climate and topography), neighbourhood measures and geometric properties, as shown in figure 4.1. These geographical factors do not deterministically predict fast or slow language change, but they favour or disfavour contact and the effectiveness of contact among groups which is associated with fast and slow language change respectively.

This chapter gave an overview of the research design of this thesis and the classes of geographical factors included in the data analysis process. The following chapter introduces the data used in this thesis, namely the transition rate data and the linguistic database used for the calculation of the language family areas, for which the geographical factors will be computed.

Chapter 5

Data

In this thesis two linguistic datasets are used for different purposes. The first dataset combines data from several databases and is used for the estimation of language change. The Glottolog database (Hammarström et al., 2016) is the second dataset and is used for the creation of language and language family areas. Further data used for the calculation of the different geographical factors will be mentioned in the methodological description in chapter 6. The estimation of language change is performed by Prof. Dr. Balthasar Bickel (Bickel, 2016), the supervisor from the Department of Comparative Linguistics, and access to the results is provided. The first subsection not only describes the resulting data, but also the basic methods applied to estimate the transition rates.

5.1 Transition Rates

In this thesis, language change is modelled as rate of change of linguistic features within language families. This means that the transition rate of features from state A to state B and vice versa is calculated. To estimate the rates, phylogenetic methods are applied. These methods are non-trivial and their suitability is debated among scholars (Heggarty, 2006). This thesis does not aim at extending or comprehensively testing phylogenetic methods, but at investigating the resulting transition rates in a geographical context. The databases used, the methodology applied to them and the resulting transition rates used for the estimation of language change are described in the following subsections.

5.1.1 Linguistic Databases

To get information on as many structural features as possible, a combined dataset including data from AUTOTYP (Nichols et al., in prep.)s, WALS (Dryer and Haspelmath, 2013), PHOIBLE (Moran, McCloy, and Wright, 2014) and ANU (Donohue et al., 2013) is used. The AUTOTYP Genealogy and Geography Database contains information on around 400 phonological and 700 grammatical features of about 2'700 languages and language varieties worldwide. The World Atlas of Language Structures (WALS) provides information about 192 structural features of roughly 2'700 languages. PHOIBLE is a database describing

phoneme inventories and distinctive feature data for phonemes in around 1'700 languages worldwide. ANU (also called World Phonotactics Database) contains information about phonotactic restrictions of over 2'000 languages and segmental data of additional 1'700 languages. In simple words, phonotactics deals with the syllable structure of languages, i.e., which sounds can be preceded and followed by which other sounds.

In these databases linguistic features are attributed to every language. Each language has one value for each collected feature. The features can take on different values representing the specific structural property of a language. For instance, in WALS, there are two to 28 possible values per feature (Dryer and Haspelmath, 2013). Here, however, only binary variables are incorporated into the computation. Ideally, languages have values for all features, but as the vast majority of languages of the world is not documented appropriately (McGregor, 2015), this is not the case for many of them. Linguistic databases generally include genealogical information of the languages and some also provide the geographical coordinates of the centres of language areas.

5.1.2 Methodology

This section outlines the applied method to estimate the transition rates of linguistic features. The following descriptions are based on the script by Bickel (2016).

5.1.2.1 Preprocessing

To be able to derive transition rates of linguistic features for each language family, it is necessary to first determine the relations of languages within a family using genealogical trees. Dediu's (2015) forests, a repository providing different taxonomies, are used for that (Dediu, 2015). Three different trees are selected for the computation: two trees with differing branch lengths are selected from Glottolog. One of them assumes uniform lengths between all nodes, while the other has branch lengths based on lexical distances based on the Automated Similarity Judgment Program (ASJP) database (Wichmann et al., 2015). The third tree used is from AUTOTYP (Bickel and Nichols, in prep.). Its branch lengths are based on structural distances. For the second and third tree, branch lengths are mapped on the trees using a genetic algorithm by Scrucca (2013).

Each tree is then matched with each feature. The trees in Dediu's databases do not necessarily end in tips (leaves of the tree) coinciding with the IDs in Glottolog. Vice versa, IDs of Glottolog are not limited to tips, but they can be found at various taxonomic levels. Thus, some adjustments are required: tips lacking data are filled with the ID of the next higher node, which is justified by the close relationship of the tip and the node. Subsequently, the trees are pruned to only retain tips with data.

5.1.2.2 Fitting CTMC Models for the Estimation of Transition Rates

For the estimation of transition rates, continuous-time Markov-Chain (CTMC) models are fitted to the trees. To do that, Dediu's forests require calibration in time (Widmer et al., 2016). As this would require extensive simulations and validation studies it is done on the basis of qualitative assumptions instead. For trees with uniform branch lengths, 1'000 years are assumed between each node. The length estimates for the lexicon-based and structure-based trees are given in Dediu's (2015) database. To result in relatively realistic times, these estimates are multiplied by five.

To get stable transition rate estimates for a feature within a language family, the transition rate is only estimated if at least ten members of that language family contain a value for this feature. To the trees fulfilling this condition, CTMC models are fitted using the package `BayesTraitsV2`. This package performs analyses of trait (here: linguistic feature) evolutions among groups of species (here: languages of a family) with available phylogenies (here: family tree) (Meade and Pagel, 2014). CTMC modelling allows a trait to change from a given state to any other state at any time. Transition rates of the trait and the likelihood, that is associated with the different states a trait can adopt, are estimated by traversing the tree (Meade and Pagel, 2014).

To facilitate rate estimates and summary statistics, only binary data is selected for the calculation, i.e. features that can adopt two states. The transition matrix in figure 5.1 is an example of a model of such a feature. Each feature has two transition rates; one rate describes the transition from state A to state B (q_{AB}) and the other the transition from state B to A (q_{BA}). For every linguistic feature in a family, two CTMC models are fitted. One model assumes equal rates (ER), i.e. q_{AB} equals q_{BA}. The second model (ARD) assumes different rates, i.e. q_{AB} and q_{BA} are unequal, allowing one transition to be more probable than the other. Then, the best fitting model for each feature in a family is chosen. CTMCs can only be fitted if feature values in a tree are non-uniform. In case a feature does not change, maximum stability is assumed indicating that q_{AB} and q_{BA} approximate zero and the equal rate model fits best.

State	A	B
A	-	q_{AB}
B	q_{BA}	-

Figure 5.1: Transition matrix of a binary feature with states A and B.

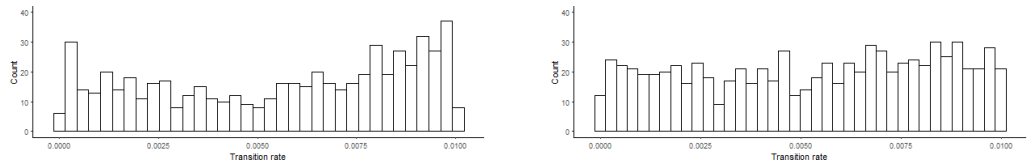
5.1.3 Data Filtering

The results contain between two and six transition rate values for each linguistic feature in each family. This is because both directions of change (q_{AB} and q_{BA}) are calculated and if available, three different trees were used to estimate the rates. For this thesis, only the highest of these rate values is selected, representing the maximum rate of change of a linguistic feature within a language family. These rates are later analysed in a geographical context. For some features in several families, transition rates are zero, indicating no change. These values are removed because they are assumed, not estimated.

5.1.4 Final Dataset

After the filtering, 1'379 transition rates of 225 different linguistic features remain. 177 of the features are grammatical and 48 are phonological. Despite that transition rates were estimated for fewer phonological features than grammatical features, they could be calculated for more families because phonological information is collected in more languages: 588 rates were estimated for 48 phonological features and 791 rates for 177 grammatical features. Thus, for some grammatical features, transition rates could only be estimated for a few language families. In terms of language families, transition rates could be estimated for 47 families, for the remaining families not enough data has been collected.

The transition rate is the first derivative of the function that describes the transition probability from one state into another state over time across the transition matrix. The values of the estimated rates range between 0.000003 and 0.01. Figure 5.2 shows the histograms of the transition rates of phonological and grammatcial features. The transition rates follow a uniform distribution and on average, the transition rates of phonological features (q=0.0056) are slightly higher than the transition rates of grammatical features (q=0.0053).



(a) Phonology **(b)** Grammar

Figure 5.2: Histograms of transition rates.

5.2 Glottolog

As was done for the transition rate data, the geographical factors are also computed on the level of language families. A dataset describing language or language family areas does not exist on a global scale, hence, they have to be assessed based on language coordinate data. This means that the coordinates of languages and their language family affiliation are required. This information is taken from the Glottolog database (Hammarström et al., 2016). This database is structured in two files: One file contains 22'924 languoids, that is, dialects, languages and several levels of genetic groupings. There are 8'397 entries for the category of language. Geographical coordinates in WGS84 are available for most of them. The resolution of the coordinates varies significantly, ranging from the metre range up to around 100 km. The other file contains the phylogenetic tree of every language family. This classification contains 242 top-level families and 188 isolates. The files are linked to determine the language family of every language. Figure 5.3 depicts all languages coloured according to their language family. The clustering of languages and language families in the tropics is clearly visible (as described in chapter 2). One has to keep in mind that the

map is projected, thereby the trend is visually enhanced because polar regions are depicted larger than equatorial regions.

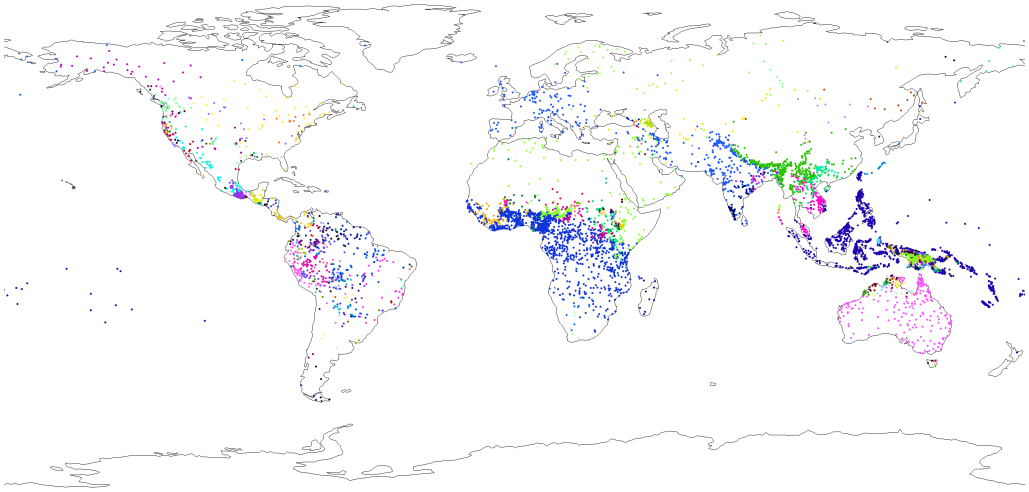


Figure 5.3: Languages of the Glottolog database (Hammarström et al., 2016). The languages are coloured according to their language family affiliation. *Data: Natural Earth, Glottolog.*

An important aspect of working with language families is their inequality in cardinal size, areal size, speaker population, etc. The difference of their cardinal size is illustrated in figure 5.4. There are only a few language families that contain more than 100 languages, the majority contains only a few languages.

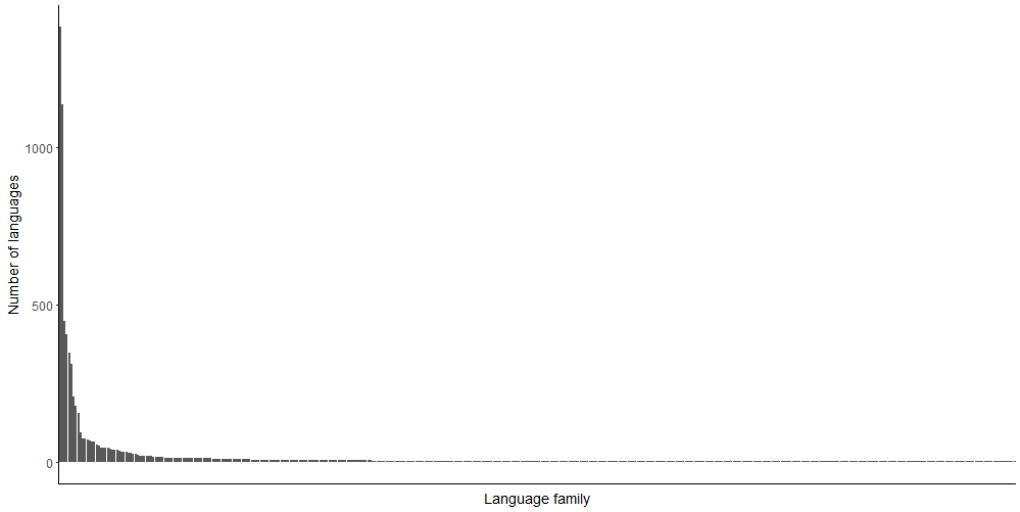


Figure 5.4: Number of languages per language family in descending order.

Chapter 6

Methodology

This chapter describes the methodology of the spatial modelling applied in this thesis (see workflow figure 4.1). The first subsection describes data filtering steps in which several languages and language families are deleted. The second subsection outlines the computation of language family areas based on the Glottolog data described above. These areas serve as a basis for the subsequent computation of the geographical factors, as they are calculated for each language family area. These values are then compared to the transition rate data in section 7. The processing, visualisations and also the subsequent analyses are realised in *R* (R Core Team, 2015), only for a few processing steps *ArcMap 10.4.1* is used.

6.1 Data Filtering

The goal of this data filtering is that only spoken languages with native speakers that have coordinates and lie on the landmass remain. Additionally, languages largely influenced by colonialism and/or globalisation are excluded. The different filtering steps lead to a deletion of 1445 languages and 11 language families. An overview of the filtering steps is provided in table 6.1 and a more detailed description of them is given in the following.

Table 6.1: Filtering steps and the number of languages that are deleted. In total, 1445 languages are removed.

Filtering step	**Number of languages removed**
Missing coordinates	767
Non-genealogical languages (pseudo-families)	169
Spurious, unclassified and ancient languages	291
Colonial languages	71
Inaccurate coordinates	147

Due to missing coordinates, 767 of the 8395 languages are deleted. In doing so, the language family South Omotic containing 5 languages and the isolate Yurumanguí are deleted. Further, five Pseudo-Families are deleted. Some languages are treated as families in the database although they are non-genealogical. 125 sign languages are deleted because speech

is seen as the primary medium of human languages (McGregor, 2015). Contact languages such as pidgins and further mixed languages deleted because they are mainly used as *lingua francas*. This lead to a deletion of 38 languages. Six more languages that are not considered as the main means of communication in a society are excluded as well. These are members of the pseudo-families Speech Registers and Artificial Languages.

Moreover, spurious, unclassified and ancient languages are removed as well: spurious languages are languages mentioned in literature, but their existence is doubted (Hammarström et al., 2016). 221 languages belonging to this 'family' are deleted. There are 64 languages of which the existence is proven, but they cannot be classified genetically (Hammarström et al., 2016). They are deleted in order not to influence the creation of the language family areas. Furthermore, six ancient languages are deleted because determining language areas for languages that do not exist anymore is not aimed at here.

Some more languages are intentionally omitted because, having resulted from colonialism or globalisation, they are much younger than the languages of interest. Hence, the inclusion of such languages would produce non-comparable results and time-depth confusion. 71 languages are identified as 'colonial languages', amongst others, Afrikaans, English Creoles and several creoles based on languages such as Portuguese and Spanish.

Languages with inaccurate coordinates are also deleted. Intersecting the language points with the landmass leads to a deletion of 147 languages. For the intersection, the 10m resolution Natural Earth dataset by ESRI (2014) is used. This high resolution preserves more languages than the world maps with resolutions of 50 or 100 meters. This may be due to the fact that some coordinates are of very high resolution. It is assumed that coordinates of language points located in water areas are erroneous and are thus deleted. The vast majority of the removed languages are located in the region of Papunesia, an area with a lot of islands. Most likely, these coordinates were not set with enough accuracy which, in the case of small islands, results in the points lying in the ocean. A further reason could be that world maps of different sources vary. Hence, setting the coordinates on a different map than the one used for the intersection could result in some points being placed in water areas although their coordinates were set accurately. In this filtering step, languages of 22 families are deleted. 108 of them are Austronesian and ten belong to the Indo-European languages. In other families, the maximum number of languages deleted is three. One of the 22 families is the isolate Chono leading to the deletion of this one-member family.

The final dataset contains 6'950 languages, which belong to 419 different language families. 186 of these are language isolates.

6.2 Modelling Language and Language Family Areas

This section describes the calculation of the language and language family polygons, as this is not a trivial process. Glottolog provides a point dataset out of which the area of a

point group has to be determined. There is no standard approach in linguistic literature for deriving language areas or language family areas from language points. It has been decided to calculate the respective Voronoi polygon for each language. These areas are then merged to language families. This decision is crucial for this thesis, because the vast majority of the subsequent calculations is based on the language and language family areas. In the following, the decision is discussed and the implementation of the method and its limitations are explained. Finally, the results are presented and it is described how they are matched to the transition rate data.

6.2.1 Point to Polygon Conversion

In GIScience the point to polygon conversion is a common issue and there are several approaches to tackle this problem. Besides the most primitive method, the bounding box, there are several other approaches, e.g., the convex and concave hull (for a description see e.g. Cormen et al. (2009)) and alpha shapes or the characteristic shape (for a description see Duckham et al. (2008)). When considering the points as languages or centres of languages, these methods have several drawbacks. With all methods, some of the language points would lie on the border of the resulting language family area. This would not represent the points correctly as language centres. Additionally, these methods would result in contiguous language family areas, which does not necessarily reflect reality. Moreover, especially the convex and concave hulls are sensitive to outliers, i.e. a language family area would be greatly distorted if one language centre lies far apart from the other languages belonging to the same language family. In order to circumvent these drawbacks, the Voronoi method is used to calculate the language family areas. This method was also used by Hammarström and Güldemann (2014) to derive language polygons.

6.2.1.1 Voronoi Method

Given a plane with some points, the Voronoi method divides the area into polygons based on the nearest-neighbour rule, i.e. a point is associated with the nearest part of the plane. This results in a polygonal partition of the whole area based on the perpendicular bisectors between pairs of points (Aurenhammer, 1991). Figure 6.1 shows points in a plane and their associated Voronoi polygons. In case of language points, the interim result is the division of an area into language polygons. To obtain language family polygons, the language polygons of the languages belonging to the same family are merged. If the family members are not adjacent

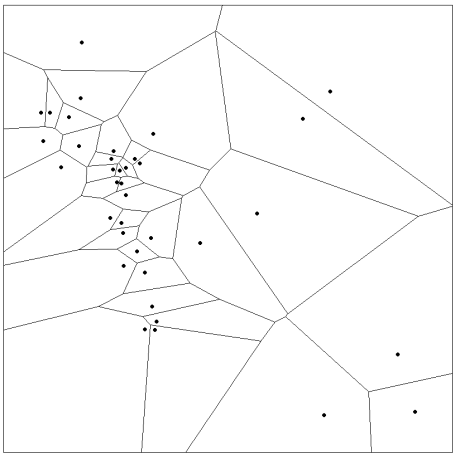


Figure 6.1: Voronoi polygons of a set of points in a plane.

(e.g. if there is an outlier), the merging results in multipart polygons. This step is explained and illustrated below.

Clear advantages of Voronoi polygons compared to the abovementioned methods are that the points represent the language centre more precisely and outliers do not distort the area, but are represented as separate polygons. However, the Voronoi method also has some disadvantages. For instance, Voronoi polygons do not represent multilingual areas, whereby the hulls could represent overlaps of different language families. A further disadvantage is the polygonal partition of the whole landmass area. As a result, uninhabited regions of the world are also attributed to languages.

6.2.1.2 Representation of Languages

A general problem of language area determination, based on point data of linguistic databases, is that all languages are represented in the same way, i.e. languages with a big speaker population cannot be distinguished from languages with only a few speakers. This information would be important for a more precise determination of the language areas. In general, a big speaker population suggests a rather big language area (Bromham et al., 2015). If this information was available, the Voronoi polygons could be weighted, resulting in relatively bigger language areas for languages with a larger speaker population.

6.2.2 Implementation

In order to calculate the language family areas, the coordinates of the language points and their family affiliation is needed. Some languages, however, have to be relocated because their coordinates are not unique. Duplicates are problematic for applying the Voronoi method, because only one of the duplicates can be represented as a polygon. The relocation of duplicates is described in the following subsection.

6.2.2.1 Relocation of Duplicates

43 coordinate pairs of languages are non-unique. Most of these (25) have one duplicate and some (11) have two. Four times there are four languages at the same location, twice five languages and once there are 10 languages with the same coordinate pair. Theoretically it is possible that some languages have the same centre, but especially in the case of multiple duplicates it is more likely that the coordinates were not set with a high accuracy.

For every group of duplicates with the same coordinate pair, one of the duplicates (languages) stays in this position, and the other languages are relocated. For this, random bearings and random distances between 3 and 8 km are used. Random parameters are used to avoid an artificial circle-like arrangement of the relocated languages. The range of distance is chosen because the resolution of the point grid used is 5 km at the equator (see description below). Thus, shifting the points between 3 and 8 km should result in a representation of these points without distorting the data too much in areas with high language density. Figure 6.2 shows the example where 10 languages have the same coordinates, namely the coordinates of red point with the black border. After the relocation, one language remains at this location, the other nine points are scattered around the original coordinates. After this relocation of duplicates, the data is suitable for the application of the Voronoi method.

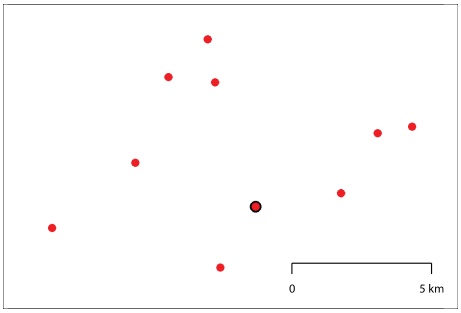


Figure 6.2: Relocation of 9 duplicates (red points) within a distance of 8 km from their original location (point with black border).

6.2.2.2 Voronoi Calculation by IDW

There is no straightforward implementation of spherical Voronoi polygons in R, because existing functions from the packages `deldir` or `dismo` only handle planar data. Thus, the Inverse Distance Weighting (IDW) function `idw` of the package `gstat` (available at https://github.com/ edzer/gstat/) is used as a workaround, because this function is able to handle spherical data. IDW is a method for multivariate interpolation based on a given set of points. Values of unknown points are calculated using a weighted average of the values at the existing points. For a more detailed description of IDW, see Isaaks and Srivastava (1989).

The procedure is illustrated in figure 6.3 showing southern India and Sri Lanka. Using IDW, the known points (i.e. the languages) and the locations where the data will be estimated have to be defined. The latter are represented by a regular point grid, i.e. a regularly distributed set of points with a certain number of rows and columns. Thus, it is not a spherical point grid in which the points are distributed regularly across the globe, but it is planar. This is necessary for the subsequent rasterization of the output point grid. The equatorial resolution of the point grid is approximately 5 km with increasing resolution towards the polar regions. Figure 6.3a shows the language points in different colours. The point grid is illustrated as small black points. As the input data is nominal (languages), no interpolation is made, but to each location of the output point grid the ID of the nearest language is assigned. Thus, the IDW output is a point grid of which each point is assigned to a language (see coloured point grid in figure 6.3b). The point grid is then rasterized, resulting in a raster of which the cell values represent the language areas (see figure 6.3c). In order to obtain

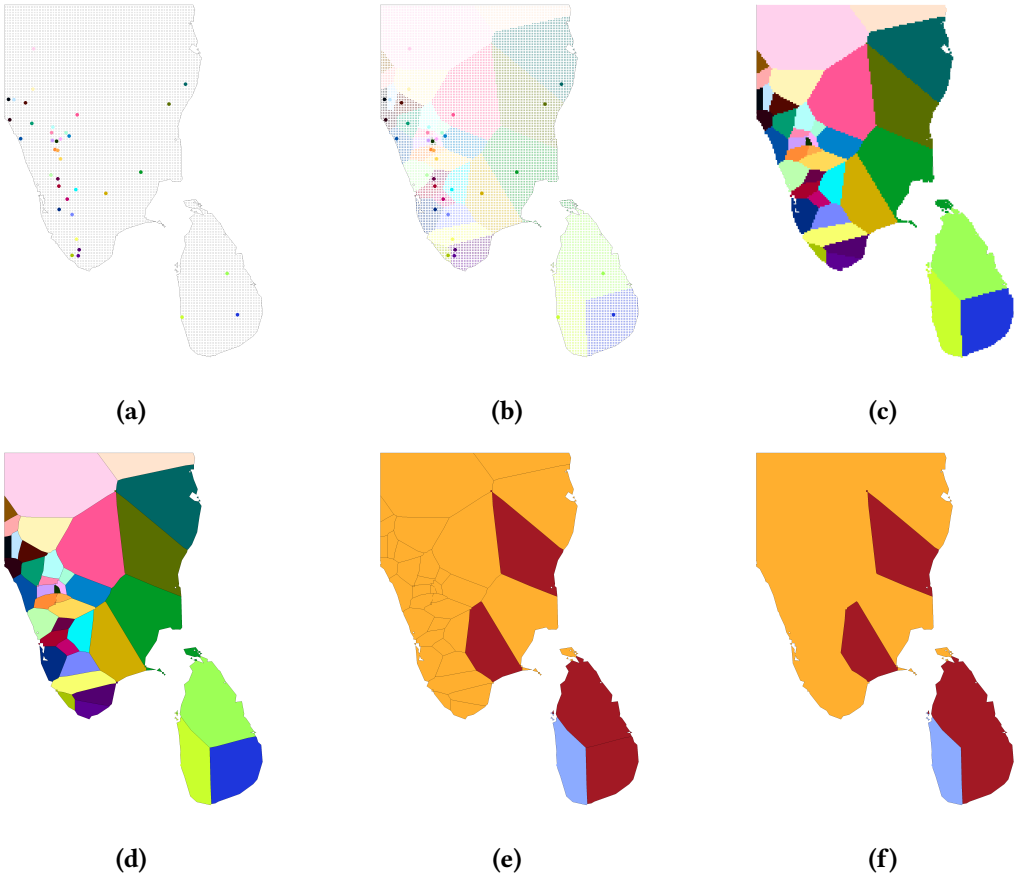


Figure 6.3: Implementation of the Voronoi method.(a) shows the initial situation: languages are represented as coloured points and the point grid is depicted in black. (b) shows the assignment of the point grid to the nearest language. (c) shows the rasterisation and (d) the raster to polygon conversion resulting in language polygons. In (e) the language polygons are coloured by their family affiliation and (f) shows the merged language family polygons. *Data: Natural Earth, Glottolog.*

a polygon dataset, the raster is converted to polygons. Raster cells with similar values are merged to one polygon, representing the area of one language (see figure 6.3d). The raster to polygon conversion is performed in ArcMap using the option *simplify polygons* which omits the cell structure of the polygons. This is done to diminish the different latitudinal resolution when e.g. calculating the perimeter of a polygon.

As described above, the language polygons are then merged based on their language family affiliation, resulting in (eventually discontiguous) language family polygons. In figure 6.3e the language polygons are coloured by language family and figure 6.3f shows the merged language family polygons.

6.2.2.3 Area Definition for Applying the Voronoi Method

The Voronoi method is applied to those landmass polygons that contain at least one language. This avoids the assignment of polygons that do not contain languages at all, mainly because no human beings live there. The global area of the total landmass is 147'049'000 km^2 and 129'987'700 km^2 if only landmasses with language points are taken into account. This difference of 17'061'300 km^2 in landmass consists mainly of Antarctica, Greenland, and the northern American Islands that account for 15'511'600 km^2. The deletion of this landmass is seen as appropriate because most of these areas are not populated. Small islands which do not contain a language are also deleted, because the language centre may be located on a nearby island or on the mainland. An example for that is illustrated in figure 6.4 showing the islands between the mainland of Denmark and Sweden. The languages are located on the mainland and thus, these polygons are deleted.

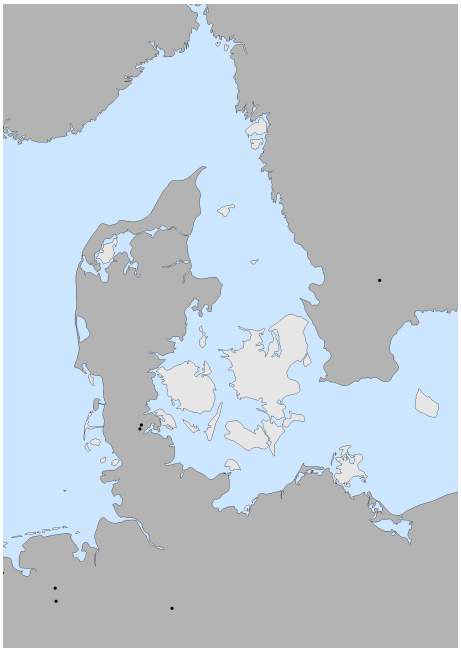


Figure 6.4: Landmass polygons that do not contain languages (black points) are removed (coloured in light grey). *Data: Natural Earth, Glottolog.*

The method is run over all landmass polygons at once. As a consequence, language points on one landmass polygon influence the polygonal partition of another landmass polygon. This method is seen as more consistent than running it over every polygon separately. For instance, the most northern part of Sri Lanka is assigned to the language Tamil (dark-green in figure 6.3b), which is centred in India. Defining the Voronoi polygons for each landmass polygon separately would result in a less correct version, namely that Tamil is only spoken in India.

6.2.2.4 Problems of the Implementation

A problem of the pseudo-spherical implementation is that the density of languages is highest in the equatorial region where the resolution of the point grid is the coarsest leading to a non-adequate representation. The resolution is not high enough to assign several cells (or points of the regular point grid respectively) to one language point. This results in rectangular language polygons (see figure 6.5).

Languages that are located very close to each other cannot be represented at all, leading to an exclusion of eleven languages in total. Eight of the excluded languages belong to the family Austronesian, and one language each to the families Madang, Tai-Kadai and Nuclear Torricelli, respectively. A closer look reveals that nine of the eleven languages lie within the area of languages of the same family. The misrepresentation thus does not have an influence on the language family polygons. The languages Lemio (Madang) and Wab (Austronesian), however, are not contained in language polygons of the same family, but by language polygons of Madi and Gwahatike, respectively, which belong to the family Nuclear Trans New Guinea (see figure 6.5).

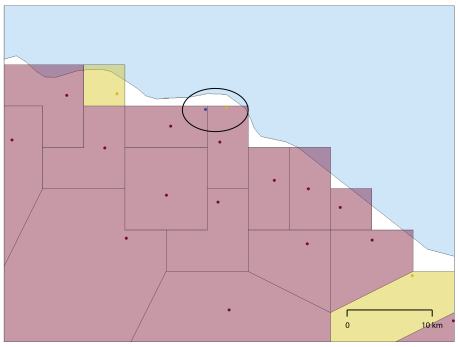


Figure 6.5: Lemio and Wab (blue and yellow point in the highlighted area) lie within language areas of languages belonging to the family Nuclear Trans New Guinea (dark red). *Data: Natural Earth, Glottolog.*

6.2.3 Resulting Language Family Areas

The result of the Voronoi method reflects the 419 language families of the Glottolog database, however, these families have to be linked to the transition rate data. As mentioned before, the number of families is hotly disputed. As a consequence, not all of the families of the transition rate data correspond to the classification by Glottolog: 31 of the 47 families of the datasets correspond, 14 families of the transition rate data are identified as one or more subfamilies in Glottolog. These subfamilies are detached from their respective families and are given top-level status. The remaining subfamilies are not altered, but the families have lost a few members. The remaining two families (Na-Dene, Macro-Ge) cannot be matched to the Glottolog data. Thus, for the further analysis, 45 families remain.

Figure 6.6 shows the resulting Voronoi polygons for all language families. The 45 language families for which geographical factors will be calculated in the following are depicted in colours. The remaining families are coloured in grey. The 45 families are distributed across the globe and all continents are represented in these families. 77.6% of the landmass area containing at least one language is covered by these 45 language families. This is quite a large fraction, however, some parts of the world are not represented very well: the southern part and east coast of South America are not part of the 45 language families and in North America there are blank spaces as well. Further, a rather large region of Sudan and Chad is not represented.

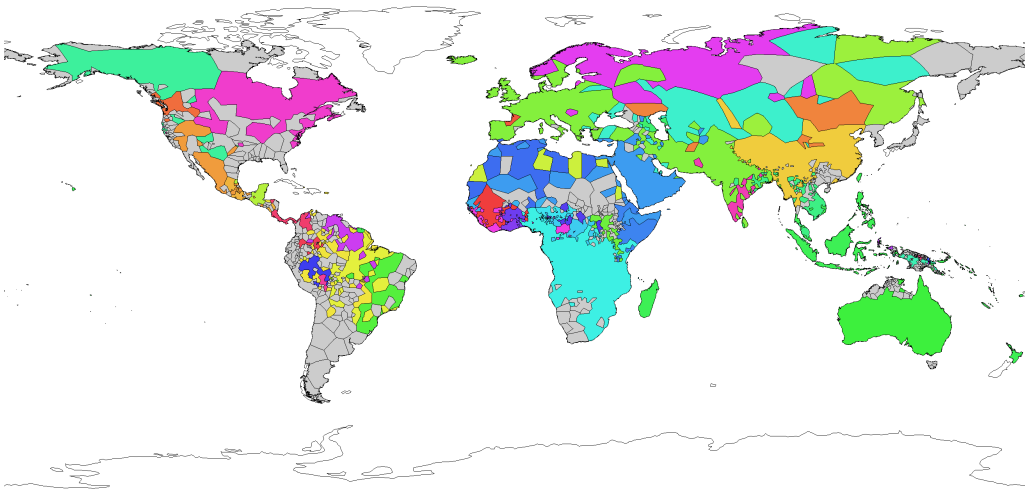


Figure 6.6: Language familiy areas for the 45 matched families are depicted in colours, the remaining families are coloured in grey. *Data: Natural Earth.*

6.3 Computation of the Geographical Factors

This section describes the calculation of the geographical factors for each language family area. In order to circumvent the drawbacks that a specific method might have, some factors are computed in various ways. An overview of all factors is provided in table 6.2. Each factor is calculated for the 45 families and the resulting values are displayed in boxplots and histograms subsequent to the respective method description. The bandwidth for the histograms is calculated by the interquartile range of the respective data divided by eight which results in an appropriate depiction of the data. All the calculations are performed spherically. Some functions in *R* provide methods that incorporate ellipsoid flattening, but in order not to mix different calculation methods, these options are not made use of. Furthermore, not all the conducted calculations are entirely spherical, for instance, when raster datasets are incorporated. This is described within the subsections.

This section follows the structure of table 6.2; first, the linguistic factor, i.e. cardinal size, is described. This is a purely linguistic factor which is incorporated because it is a dependent variable directly reflecting the divergence of languages within a family. For simplicity reasons this factor is not treated separately although it is not geographical *per se*. Second, environmental characteristics encompassing climate and topography are described. Third, neighbourhood factors are calculated based on different definitions of neighbourhood and fourth, geometric properties, mainly horizontality and compactness, are computed. After the description of the factors of each subdivision, a correlation matrix is presented to show the dependencies among the factors. This is important with regard to the subsequent interpretation of the results. The linguistic factor is included in all the matrices.

Table 6.2: Geographical factors.

Class	Description		Factor Name
Linguistic	Cardinal size		SIZE
Environmental characteristics	Precipitation	Average and standard deviation of the annual precipitation	$PREC_{AV}$ $PREC_{SD}$
	Temperature	Average and standard deviation of the maximum mean temperature	$TEMP_{MAX.AV}$ $TEMP_{MAX.SD}$
		Average and standard deviation of the minimum mean temperature	$TEMP_{MIN.AV}$ $TEMP_{MIN.SD}$
	Topography	Average and standard deviation of elevation	ALT_{AV} ALT_{SD}
		Average and standard deviation of Terrain Ruggedness Index (TRI)	TRI_{AV} TRI_{SD}
Neighbourhood measures	Adjacency	Family average of the number of adjacent languages	$ADJ_{LANG.TOT}$ $ADJ_{LANG.SF}$ $ADJ_{LANG.DF}$
		Number of adjacent families	ADJ_{FAM}
	Point distance	Family average of the number of languages within a certain distance	$PD_{100.TOT}$ $PD_{500.TOT}$ $PD_{100.SF}$ $PD_{500.SF}$ $PD_{100.DF}$ $PD_{500.DF}$
Geometric properties	Areal size		AREA
	Perimeter		PERI
	Horizontality		HOR_{MAX} HOR_{MID}
	Compactness	Area-perimeter measures	P2A IPQ
		Reference shape measures	REOCK $REOCK_{CORR}$

6.3.1 Linguistic Factor - Cardinal Size

Additionally to the geographical factors, a linguistic factor is incorporated, namely the number of languages belonging to a family. The number of points is calculated from the dataset before the Voronoi polygons were built, in order to not distort the data due to the Voronoi error.

The number of languages per family range from 1 to 1'137. This factor is normally distributed with a mean of 122 languages, but it is distorted by the four outliers that have

values above 250 (see figure 6.7). Most of the families contain between 20 and 100 languages, with Pano-Tacanan, Afro-Asiatic and Basque being the only families consisting of less than 10 languages. As mentioned in section 5 this would indicate that transition rates cannot be calculated for these families, other databases, however, contain more languages for Pano-Tacanan and Afro-Asiatic. Basque is defined as an Isolate with several dialects in Glottolog, the databases used for the calculation of the transition rates, however, define the dialects of Basque as languages making up the language family and this allows for the calculation of transition rates.

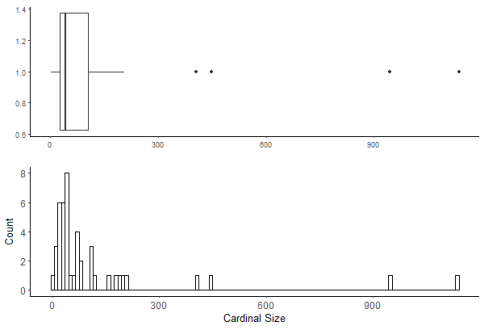


Figure 6.7: Boxplot and histogram of SIZE.

6.3.2 Environmental Characteristics

To characterisee the environment of the language families, climatic conditions and topographical characteristics of the language family areas are examined. This section describes how the different measures are calculated. For the computation of the climate factors, the paleoclimatic data model MRI-CGCM3 is used, which is generated by the Meteorological Research Institute (MRI) (2016). This data is available online (http://www.worldclim.org/paleoclimate1, accessed on 05.11.2016). The model estimates the temperature and precipitation of the Mid-Holocene around 6'000 years BP. This is more suitable than today's climatic condition because the depth of the language family trees goes back to around 10'000 years BP. The resolution of the raster data is 10 arc minutes, which corresponds to an equatorial resolution of 18.6 km. This is sufficient for the purpose of this thesis, because precipitation and especially temperature are not spatially fast changing processes.

For the calculation of the topographic characteristics ETOPO1 was used (Amante and Eakins, 2009), available on https://www.ngdc.noaa.gov/mgg/global/global.html (accessed on 02.10.2016). ETOPO1 is a global digital elevation model with a resolution of 1 arc minute, i.e. 1.85 km at the equator. There are two versions; one depicts the surface of the ice sheets of Antarctica and Greenland, the other depicts the bedrock beneath the ice. For this thesis it does not matter which version is chosen because Antarctica and Greenland were deleted.

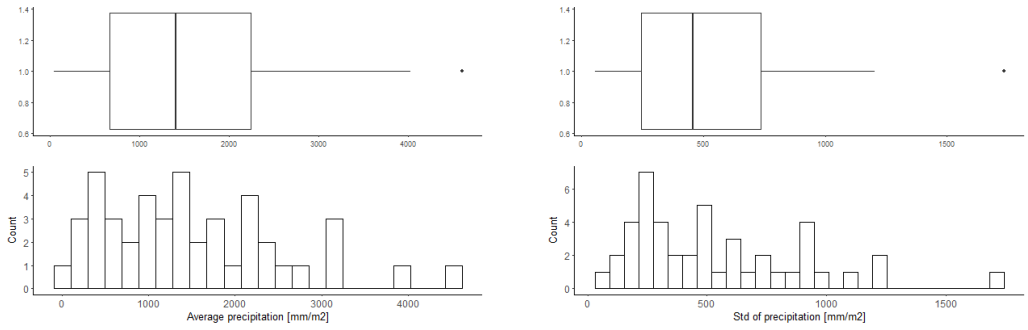
6.3.2.1 Climatic Conditions

For both the precipitation and the temperature, the average value and the standard deviation are calculated. These values stand for the average climatic condition and the climatic variability in a language family area, respectively. As the calculation of the values is based on raster data, the calculation is not entirely spherical. The grid data is converted to polygons and clipped to the areas of the language families. The areas of all polygons of every temperature or precipitation value, respectively, are added up in order to get the frequency in numbers of km^2 from which the average and standard deviation values are calculated. This is done because a simple count of pixels with similar values does not account for the data not being spherical. That is, two different values may have the same amount of pixels, but not the same area because the pixels become smaller towards the poles. Hence, the actual pixel area has to be calculated, which is done by converting the pixels into polygons.

Precipitation ($PREC_{AV/SD}$)

For assessing the precipitation in a language family area, the annual precipitation [mm/m^2] is averaged for the area of every language family ($PREC_{AV}$). The standard deviation of the precipitation values within a language family area is used as measure for the climatic variability within the respective area ($PREC_{SD}$).

The resulting $PREC_{AV}$ values follow a normal distribution with a mean of 1'502 and a rather big standard deviation of 1'060. The values range from 42 mm/m^2 to 4'603 mm/m^2. Only two families have a value above 3'500, whereby the higher value is an outlier (see figure 6.8a). Figure 6.8b shows the distribution of the $PREC_{SD}$ values ranging from 56 to 1'733. The precipitation within most of the language families varies between 200 and 500 mm/m^2 and several have a $PREC_{SD}$ value around 900.



(a) $PREC_{AV}$ **(b)** $PREC_{SD}$

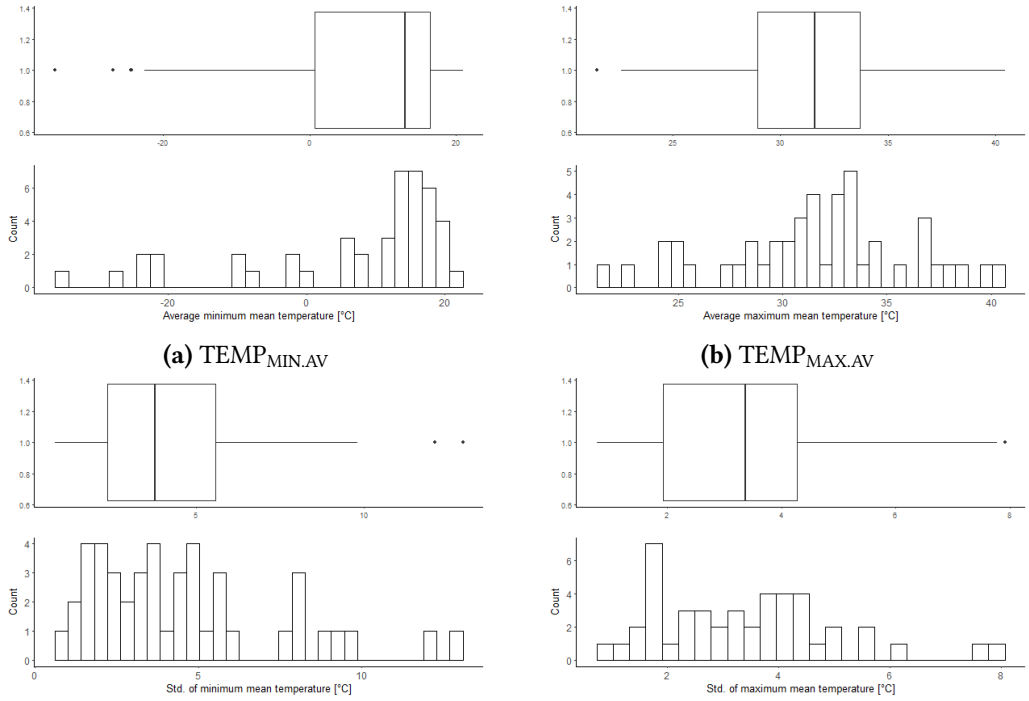
Figure 6.8: Boxplots and histograms of the precipitation measures.

Temperature ($TEMP_{MIN.AV/SD}$,$TEMP_{MAX.AV/SD}$)

The two datasets used provide the mean maximum and mean minimum temperature for every month, i.e. 12 rasters per dataset. For every pixel, the values of the month with

the highest mean maximum and the lowest mean minimum temperature are chosen. Subsequently, the average and the standard deviation of the mean maximum and the mean minimum temperature are calculated. The average values ($TEMP_{MIN.AV}$, $TEMP_{MAX.AV}$) represent a general temperature value and the standard deviation measures ($TEMP_{MIN.SD}$, $TEMP_{MAX.SD}$) are assessed as temperature variability measures within the language family areas.

Figure 6.9a depicts the result of the $TEMP_{MIN.AV}$ calculation. The values range from -35 to 21°C whereas $TEMP_{MAX.AV}$ values range from 21.5° to 40.5°C (see figure 6.9b). Most of the $TEMP_{MIN.AV}$ values lie between 10 and 20°C with three outliers in with values lower than -20°C. Most of the families have a $TEMP_{MAX.AV}$ value between 28 and 33°C. $TEMP_{MAX.AV}$ follows a normal distribution. The standard deviation values for the mean minimum temperature range from 0.76 to 13°C, whereas most values lie between 1 and 6°C (see figure 6.9c) with two outliers with values avove 10 °C. Generally, the standard deviations are rather big. The $TEMP_{MAX.SD}$ values lie between 0.77 and 7.9, whereas most values are slightly below 2 and between 2.5 and 4.5°C (see figure 6.9d). Both standard deviation values follow a normal distribution.



(a) $TEMP_{MIN.AV}$ **(b)** $TEMP_{MAX.AV}$

(c) $TEMP_{MIN.SD}$ **(d)** $TEMP_{MAX.SD}$

Figure 6.9: Boxplots and histograms of the temperature measures.

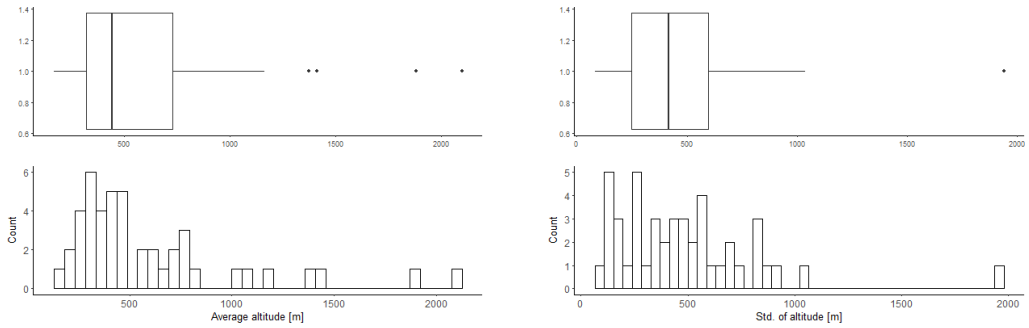
6.3.2.2 Topographical Characteristics

In this section, the average elevation, the standard deviation of the elevation values and the Terrain Ruggedness Index (TRI) are calculated for each language family area, which serve as indicators for the roughness within the area.

Altitude Measures ($ALT_{AV/SD}$)

The simplest measure is the average elevation of a language family (ALT_{AV}). However, this measure is not conclusive for language family areas containing different types of topographical landscapes, e.g. mountains and plains or when lying on a high plateau. A better measure is the standard deviation of the height values (ALT_{SD}), applied by e.g. Ascione et al. (2008) and Currie and Mace (2009). This is a primitive roughness measure, standing for the variability of altitude within an area. This method does not incorporate spatial dependencies, as it only looks at the frequency distribution of height values. The ALT_{AV} and ALT_{SD} are calculated in the same way as the climatic measures described above. The resolution of the grid was lowered to 5 arc minutes due to computational efficiency, however, based on the method applied, higher and lower resolutions result in similar values.

The ALT_{AV} values range from 165 m to 2'100 m (see figure 6.10a). Most language families have an average altitude between 200 and 500 m above sea level. Only a few languages have values above 700 m and four outliers have values above 1'300 m. Figure 6.10b shows the standard deviations of the altitude ranging from 79 m to 1'930 m, whereby the maximum value is an outlier. Most of the families have values below 600 m, several families have values between 100 and 300 m. These values follow a normal distribution.



(a) ALT_{AV} **(b)** ALT_{SD}

Figure 6.10: Boxplots and histograms of the altitude measures.

Terrain Ruggedness Index ($TRI_{AV/SD}$)

Riley, DeGloria, and Elliot (1999) have developed the Terrain Ruggedness Index (TRI) which is based on a terrain model and incorporates neighbouring cells to measure terrain roughness. The TRI value is calculated for each cell by adding up the absolute elevation difference between this cell and its eight neighbour grid cells (Riley, DeGloria, and Elliot, 1999). In figure 6.11 a hypothetical digital elevation model is shown. Adding up the absolute differences (10 (top left), 5, 0, 11, 2, 5, 2, 8 (bottom right)) results in a TRI value of 43 for the centre cell.

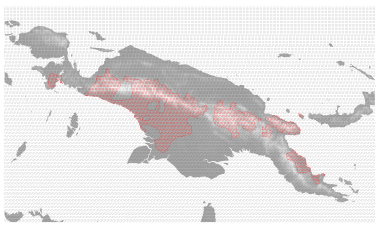
160	155	150
161	150	152
155	148	142

Figure 6.11: Nine cells of a hypothetical digital elevation model. The TRI value of the cell in the middle is 43.

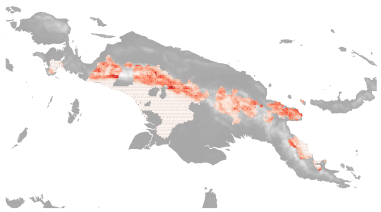
As the TRI is calculated per cell, the average value (TRI_{AV})
and the standard deviation of the TRI values (TRI_{SD}) are calculated for the areas of the language families. As mentioned before, the terrain model used has a resolution of 1 arc minute which corresponds to 1.85 km at the equator, however, in order to distinguish mountains from plains, this resolution is too high. Testing showed that a resolution of 5 arc minutes (9.28 km at the equator) is an appropriate resolution for the detection of macro-topography restricting mobility. Thus, the resolution of the terrain model is reduced to 5 arc minutes.

This method is implemented spherically, which is illustrated for the language family area of Nuclear Trans New Guinea (see figure 6.12). To get an impression of the topography of this region, figure 6.12a shows a detailed shaded relief of Papua New Guinea and the nearby islands (resolution of 1 arc minute). In order to obtain spherical data, a spherical point grid is created, i.e. a point set with spherically regularly distributed points across the globe. This point grid has a (constant) resolution of 5 arc minutes. Figure 6.12b illustrates the point grid in black which is laid over the terrain model. The points lying within the language family area are highlighted in red. Subsequently, each point is assigned the elevation value of the raster cell it lies in. Then, for each point, the TRI value is calculated by incorporating the eight nearest points. The resulting TRI values for Nuclear Trans New Guinea are depicted in figure 6.12c . The dark red points in the mountain areas represent high TRI values (up to 11'650 m), indicating a rough terrain. The light values in the plain are mostly below 100 m, indicating a flat terrain. The distinction of plain and mountain range is very clear which indicates that a resolution of 5 arc minutes is suitable for the detection of mountain ranges that restrict mobility.

(a) Detailed shaded relief.



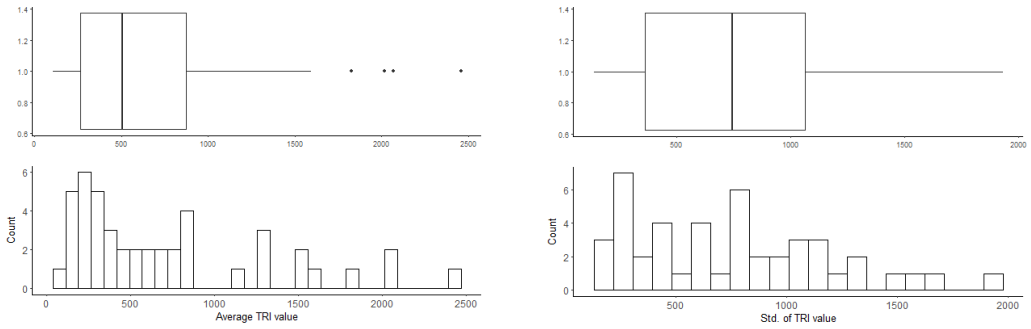
(b) Terrain model overlayed by spherical point grid.



(c) TRI values of the grid points. High values are depicted in dark red.

Figure 6.12: Illustration of the TRI calculation for the language family Nuclear Trans New Guinea (language family area marked in red). *Data: Natural Earth, Glottolog, ETOPO1.*

For all families, the resulting TRI_{AV} values range from 105m to 2'460 m (see figure 6.13a). Most of the families have values between 200 and 500 m, but the values are scattered quite widely, several families have values around 800 m and eleven families have higher values. There are four outliers with high values. Figure 6.13b shows the distribution of the TRI_{SD} values. The minimum value is 138 and the maximum 1'933 m. The values are broadly scattered, a lot of them lie between 200 and 800 m.



(a) TRI_{AV} **(b)** TRI_{SD}

Figure 6.13: Boxplots and histograms of the TRI measures.

6.3.2.3 Correlation Matrix

Except for $TEMP_{MAX.AV}$ there are strong correlations among the temperature measures and also among the topographic measures (see table 6.3). There are also some correlations between the average precipitation and temperature measures and negative dependencies between temperature and topography.

Table 6.3: Correlation matrix of the environmental factors. The correlation coefficient (Spearman's *rho*) between each pair is depicted.

Environmental Characteristics	$PREC_{AV}$	$PREC_{SD}$	$TEMP_{MIN.AV}$	$TEMP_{MIN.SD}$	$TEMP_{MAX.AV}$	$TEMP_{MAX.SD}$	ALT_{AV}	ALT_{SD}	TRI_{AV}	TRI_{SD}
SIZE	0.15	0.47	0.11	0.25	0.10	0.13	0.12	0.06	0.00	0.04
$PREC_{AV}$		0.56	0.84	-0.42	-0.12	-0.59	-0.31	-0.13	0.16	0.22
$PREC_{SD}$			0.39	0.12	-0.02	0.06	0.11	0.30	0.36	0.39
$TEMP_{MIN.AV}$				-0.57	0.26	-0.66	-0.48	-0.28	-0.09	0.04
$TEMP_{MIN.SD}$					-0.41	0.83	0.56	0.71	0.43	0.42
$TEMP_{MAX.AV}$						-0.21	-0.34	-0.51	-0.63	-0.55
$TEMP_{MAX.SD}$							0.59	0.65	0.35	0.36
ALT_{AV}								0.71	0.64	0.53
ALT_{SD}									0.77	0.84
TRI_{AV}										0.92

6.3.3 Neighbourhood Measures

In this section, different neighbourhood measures are calculated. In order to circumvent the drawbacks of a specific method, these measures are calculated in two ways, by using a different definition of neighbourhood, namely adjacency and point distance. For these measures, the language points and areas from the families for which no transition rates could be calculated are incorporated as well. The reason is that they are also neighbours of the families for which the number of neighbours is calculated. The applied methods are illustrated with the example of Pano-Tacanan, a family with seven languages spoken in northwestern South America.

6.3.3.1 Adjacency Measures (ADJ)

For these measures, neighbourhood is defined by polygon adjacency. This is done on two levels, namely on language family level and on language level. The adjacency measures are highly dependent on the calculated Voronoi polygons. The normalization of the count of neighbours by perimeter and area of the language or language family polygon was not done because small families (usually having a lot of neighbours) get high values and rather big families are underestimated. Thus, it was decided to only calculate the absolute number of neighbours.

Average of Languages (ADJ_{LANG})

For each language in a language family, the number of neighbouring languages in total ($ADJ_{LANG.TOT}$) is calculated. Furthermore, it is differentiated between neighbours of the same family ($ADJ_{LANG.SF}$) and of different families ($ADJ_{LANG.DF}$). Then the average of the languages is calculated to get the value for the respective language family. The calculation of a ratio of $ADJ_{LANG.DF}$ and $ADJ_{LANG.SF}$ is not possible because for some languages, one of the values is zero. As the Voronoi polygons do not represent 11 languages (see explanation in chapter 6.2), no value is calculated for these missing languages. Eight of the missing languages belong to Austronesian, a language family with 1137 languages. Therefore, this error does not have a strong impact on the resulting average values for Austronesian. Figure 6.14 shows the language areas of Pano-Tacanan in red and the language areas of other families in grey. Based on adjacency, four of the members

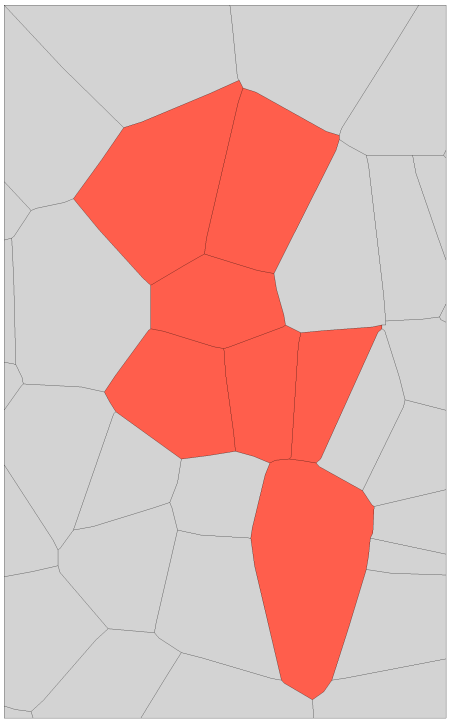


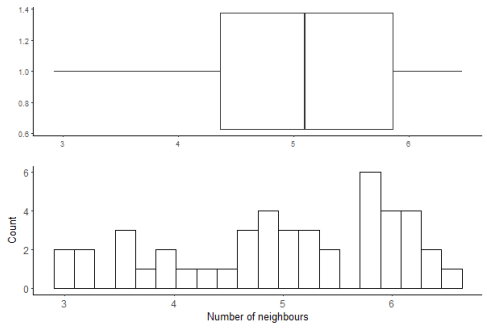
Figure 6.14: Languages of the family Pano-Tacanan are coloured in red, while languages belonging to other families are depicted in grey.

have six neighbours, two have five neighbours and one language has eleven neighbours (see table 6.4). The languages of the family thus have 6.43 neighbours on average ($ADJ_{LANG.TOT}$). $ADJ_{LANG.SF}$ and $ADJ_{LANG.DF}$ are calculated in the same way, resulting in 2.57 and 3.86 neighbours, respectively.

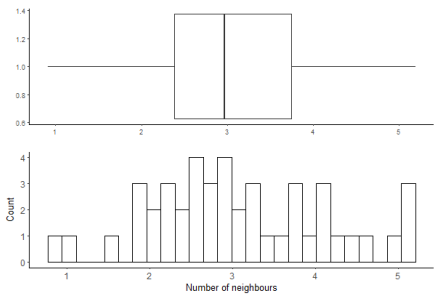
Table 6.4: Calculation of the values of the ADJ_{LANG} for the language family Pano-Tacanan. The seven member languages are depicted in the rows and the number of adjacent neighbours in total (TOT), from the same family (SF) and from a different family (DF) in the columns. The average is the value of the respective measure.

ADJ_{LANG}	Language 1	Language 2	Language 3	Language 4	Language 5	Language 6	Language 7	Average
TOT	6	5	6	6	6	5	11	6.43
SF	2	2	4	2	4	2	2	2.57
DF	4	3	2	4	2	3	9	3.86

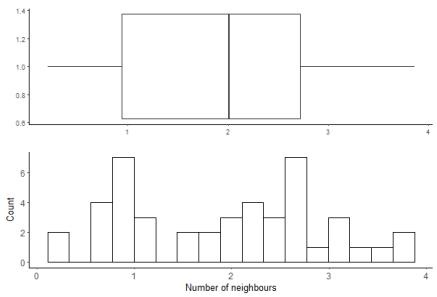
The resulting average values of $ADJ_{LANG.TOT}$ lie between 2.9 and 6.5 (see figure 6.15a). The majority of language families have between 4.5 and 6 neighbours on average. Language families have $ADJ_{LANG.SF}$ values between 0.9 and 5.2 and $ADJ_{LANG.DF}$ values between 0.2 and 3.86 (see figures 6.15b and 6.15c). The $ADJ_{LANG.SF}$ values follow a normal distribution with a mean value of 3.08 and a standard deviation value of 1.06. The values of the average neighbours of different families are distributed in a bimodal manner with peaks at around 1 and 2.7.



(a) $ADJ_{LANG.TOT}$



(b) $ADJ_{LANG.SF}$



(c) $ADJ_{LANG.DF}$

Figure 6.15: Boxplots and histograms of the ADJ_{LANG} measures.

Language Family (ADJ_{FAM})

The second level of calculation is the family level and thus the language family polygons are analysed. For every language family, the number of adjacent language families is counted. The example of Pano-Tacanan is illustrated in figure 6.16. Pano-Tacanan (red) is surrounded by ten different language families.

The resulting number of neighbouring families range from 1 to 60 (see figure 6.17), showing the two extremes. Most of the families have between 10 and 15 neighbours and there are four outliers with more than 40 neighbours.

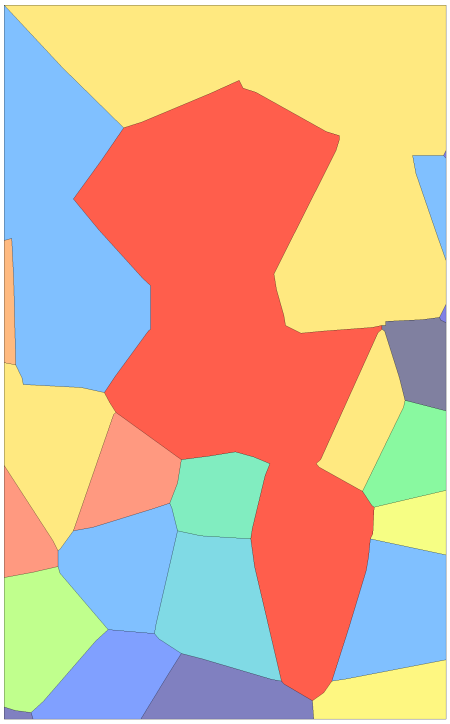


Figure 6.16: The language family area of Pano-Tacanan is depicted in red. It is surrounded by 10 language families.

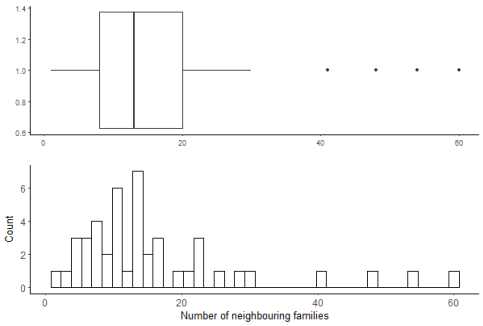


Figure 6.17: Boxplot and histogram of ADJ_{FAM}.

6.3.3.2 Point Distance Measures (PD)

Adjacency is highly dependent on the Voronoi polygons. Therefore, an independent measure using point distance is also calculated. The measure is based on the coordinates of the initial language points. Neighbourhood is defined as a specified inter-point distance, resulting in language density values in circles surrounding a specific language, similar to the work by Köhli (2013). For every language, the languages lying within a certain distance are counted and subsequently an average value per language family is built, similar to the ADJ_{LANG} measures.

This is done for close neighbours lying within 100 km ($PD_{100.TOT}$) and for more distant neighbours lying within 500 km ($PD_{500.TOT}$). These distances are defined after having tested different values between 25 and 800 km, whereby most of them correlate highly. Moreover, for every inter-point distance it is distinguished between the neighbouring languages belonging to the same or to a different language family, resulting in the following additional measures: $PD_{100.SF}$, $PD_{100.DF}$, $PD_{500.SF}$, and $PD_{500.DF}$. Similar to the ADJ_{LANG} measures, the calculation of the ratio of neighbours of the same and of different language families is not possible, due to the fact that for some languages, one of the two values equals zero.

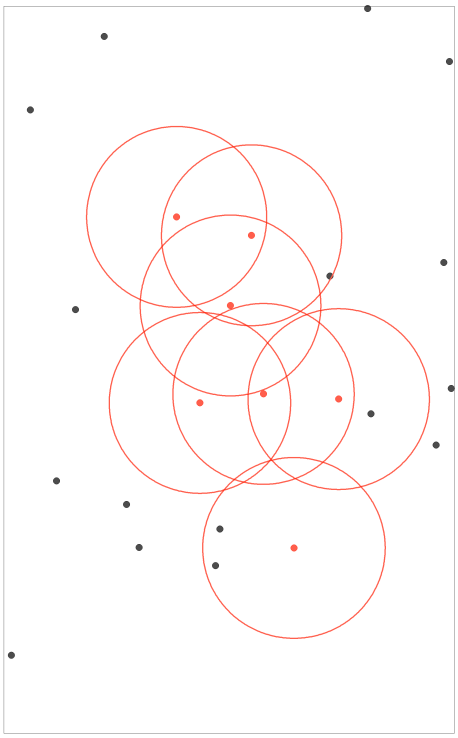


Figure 6.18: The seven languages of Pano-Tacanan and the circle of 100 km around them are depiced in red. Languages belonging to different families are illustrated as black points. *Data: Glottolog.*

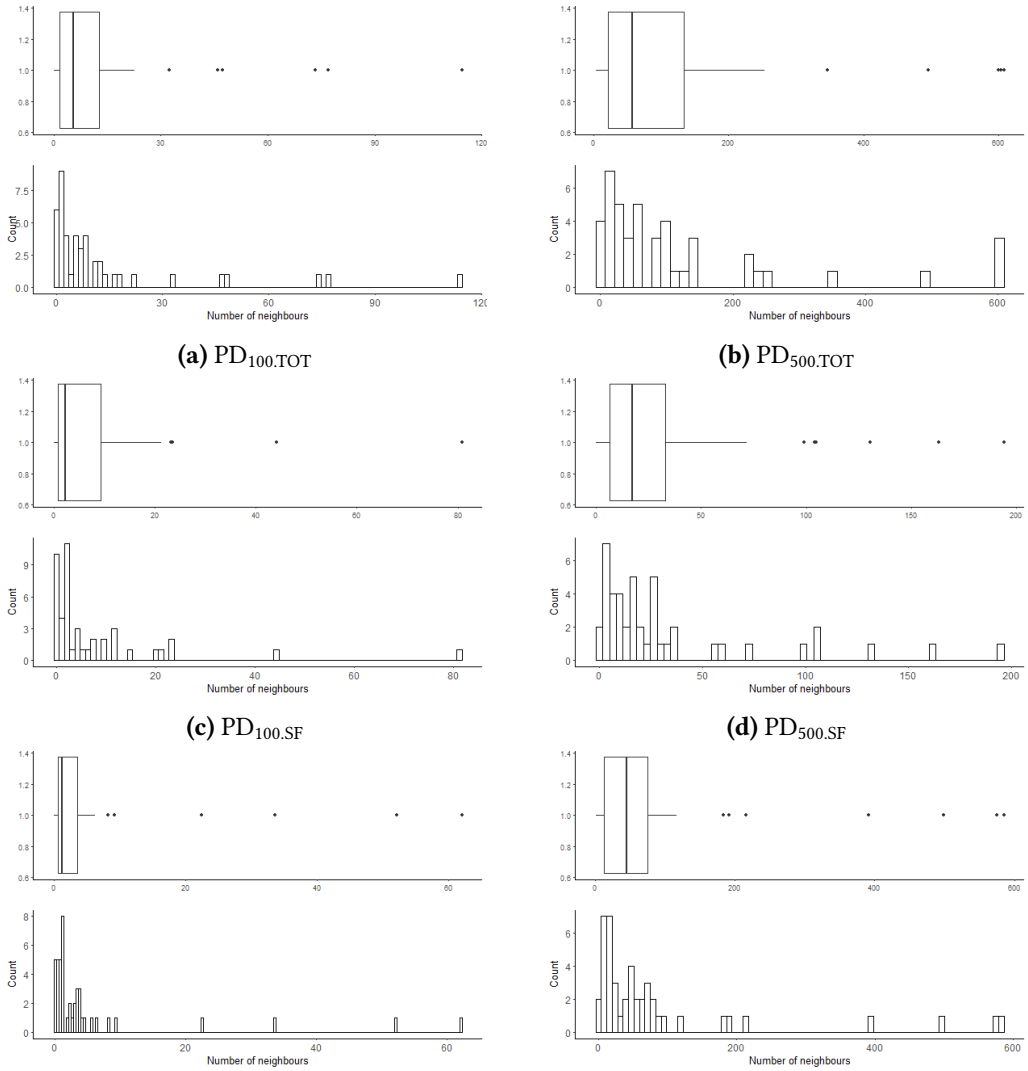
Figure 6.18 illustrates the example of Pano-Tacanan for the neighbours within 100 km. Pano-Tacanan languages are coloured in red and circles with a radius of 100 km are displayed. Table 6.5 depicts the number of languages within this distance for the seven member languages, also differentiating between neighbours of the same and of different families. The languages have 1.71 neighbours on average ($PD_{100.TOT}$). $PD_{100.SF}$ and $PD_{100.DF}$ are calculated in the same way, resulting in 1.14 and 0.57 neighbours on average, respectively.

Table 6.5: Calculation of the values of PD_{100} for the language family Pano-Tacanan. The seven member languages are depicted in the rows and the number of adjacent neighbours in total (TOT), from the same family (SF) and from a different family (DF) in the columns. The average is the value of the respective measure.

PD_{100}	Language 1	Language 2	Language 3	Language 4	Language 5	Language 6	Language 7	Average
TOT	1	3	1	1	2	2	2	1.71
SF	1	2	1	1	2	1	0	1.14
DF	0	1	0	0	0	1	2	0.57

Figure 6.19 shows the histograms and boxplots of the different PD measures. The minimum values for $PD_{100.TOT}$ and thus also for $PD_{100.SF}$ and $PD_{100.DF}$ are zero. Within 100 km, the languages of a family have at most 114 neighbours in total, 81 neighbours of the same family and 62 neighbours of a different family. Within 500 km, languages have between 4 and 608 neighbours on average . They may not have neighbours from the same family, but at least 1 neighbour from a different family. The maximum number of neighbours

belonging to the same family within 500 km is 194 and from a different family this value is 585. The PD_{100} measures show peaks relatively close to zero and some outliers above 30 ($PD_{100.TOT}$) and above 20 ($PD_{100.SF}$, $PD_{100.DF}$) respectively. The distribution of the PD_{500} measures also show peaks relatively close to zero, but the number of neighbours of the families is scattered more widely.



(a) $PD_{100.TOT}$ **(b)** $PD_{500.TOT}$

(c) $PD_{100.SF}$ **(d)** $PD_{500.SF}$

(e) $PD_{100.DF}$ **(f)** $PD_{500.DF}$

Figure 6.19: Boxplots and histograms of the PD measures.

6.3.3.3 Correlation Matrix

There are strong correlations among different neighbourhood measures. For instance, there are strong relationships among the PD measures and between these and $ADJ_{LANG.SF}$ (see table 6.6). The dependencies are weaker among the adjacency measures. There are several negative relationships between measures that differentiate neighbours of the same and of different language families, but except for the relationship between $ADJ_{LANG.SF}$ and

$ADJ_{LANG.DF}$, they are not strong. There are also dependencies between the cardinal size of a family and several neighbourhood measures.

Table 6.6: Correlation matrix of the neighbourhood measures. The correlation coefficient (Spearman's *rho*) between each pair is depicted.

Neighbourhood Measures	ADJ.FAM	$ADJ_{LANG.TOT}$	$ADJ_{LANG.SF}$	$ADJ_{LANG.DF}$	PD.100	$PD_{100.SF}$	$PD_{100.DF}$	PD.500	PD.500.SF	$PD_{500.DF}$
SIZE	0.41	0.24	0.58	-0.39	0.51	0.53	0.44	0.51	0.73	0.36
ADJ_{FAM}		0.12	-0.05	0.13	0.02	-0.03	0.17	0.08	0.08	0.05
$ADJ_{LANG.TOT}$			0.50	0.42	0.55	0.48	0.62	0.72	0.49	0.73
$ADJ_{LANG.SF}$				-0.52	0.80	0.88	0.48	0.76	0.89	0.58
$ADJ_{LANG.DF}$					-0.29	-0.43	0.10	-0.09	-0.44	0.11
PD.100						0.97	0.84	0.94	0.88	0.86
$PD_{100.SF}$							0.71	0.88	0.91	0.76
$PD_{100.DF}$								0.89	0.63	0.93
PD.500									0.85	0.95
PD.500.SF										0.70

6.3.4 Geometric Properties

In this section, eight geometric measures of the language family areas are calculated. These encompass basic measures like the area and the perimeter of a polygon and more advanced measures that analyse the shape of the language family area, namely the horizontality and the compactness of a shape. Multipart polygons are treated the same way as single polygons.

6.3.4.1 Areal Size (AREA)

The areal size is the (spherical) area of the language family polygons. In case of multipart polygons, the areas of the different polygons are added up in order to get the complete areal size of a language family.

The resulting shape areas range from 6'500 km^2 to approximately 10.8 million km^2 (see figure 6.20). Two thirds of the language families have an area below 2 million km^2. The distribution is skewed to the left and the area of only a few language families exceeds 3 million km^2. There are three outliers with values higher than 8 million km^2.

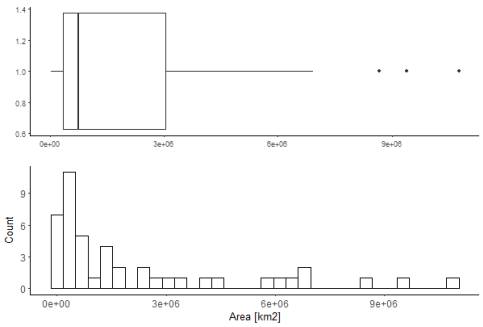


Figure 6.20: Boxplot and histogram of AREA.

6.3.4.2 Perimeter (PERI)

The perimeter is calculated for each language family polygon. In the case of multipart polygons the perimeter values of the single polygons are added up. However, the method `perimeter` from the package `geosphere` (available at: https://github.com/cran/geosphere) does not account for holes in a polygon. That is, only the outline of a polygon is calculated. There are two potential problems using the perimeter as a measure. First, the perimeter depends on the resolution of the polygons, which differs in latitudinal direction. To circumvent this problem at least partly, the option *simplify polygons* was used in the polygonisation step, as described in section 6.2 . Second, the measure may not be meaningful, for instance, in case of nearby islands that belong to the same language family. The perimeter is calculated for each island polygon although this may not be the actual border of the language family.

The resulting perimeter values range from 475 km to 94'500 km (see figure 6.21). Most of the language family areas have a perimeter between 5'000 and 20'000 km and there are three outliers exceeding 45'000 km. This distribution is also skewed to the left, similar to the distribution of the area values.

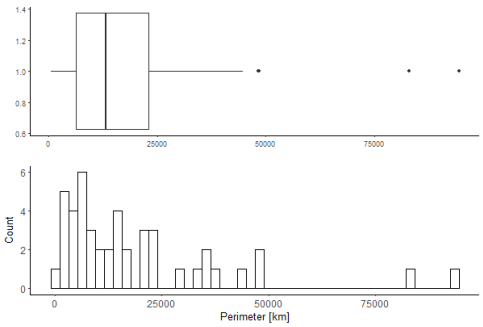


Figure 6.21: Boxplot and histogram of PERI.

6.3.4.3 Horizontality (HOR_{MAX}, HOR_{MID})

Horizontality, the ratio of the east-west to the north-south expansion of a language family, is a measure used by Hammarström (2010). He used the coordinates of the member languages

(point data) to assess the horizontality of language families. In this thesis, instead of the language points, the language family areas are used to calculate horizontality. This means that the coordinates of the bounding boxes of the language family polygons are used. The north-south expansion is defined as the distance between the most northern and the most southern latitude along a similar longitude.

The east-west expansion is calculated in two different ways: One possibility is the maximum span, i.e. the distance of the maximum longitudinal span. This is the distance between the longitudes at the latitude closest to the equator or at the equator itself, in case a language family is spread on both hemispheres. This method is called maximum horizontality (HOR_{MAX}). Figure 6.22 illustrates this on a hypothetical polygon: the bounding box is visualized in black and the maximum longitudinal span is depicted in red. The second method, which is used by Hammarström (2010), calculates the distance between the most eastern and western longitude at the latitude lying in the middle of the most northern and most southern latitude. This method is referred to as HOR_{MID}. The blue line in figure 6.22 depicts the span used for this calculation. The bigger the north-south expansion of a family, the bigger the difference between the two measures, unless the family is equally spread on both hemispheres.

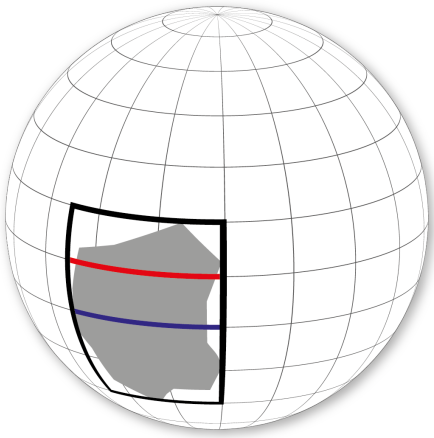
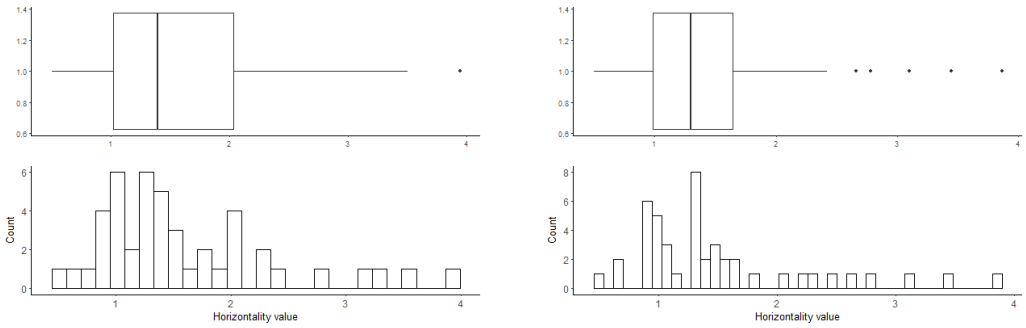


Figure 6.22: Illustration of the calculations of the east-west spread for a hypothetical polygon. The maximum spread is depicted in red, the mid spread in blue.

The HOR_{MAX} values range from 0.5 to 3.9 (see figure 6.23a). The values for most families lie between 0.8 and 1.5, i.e. their east-west and their north-south expansion are about the same. The distribution shows another peak at approximately 2, indicating that several families are of horizontal shape. Figure 6.23b shows the distribution of the resulting HOR_{MID} values ranging from 0.496 to 3.866. This distribution shows two peaks, the first is at approximately 1, the second at around 1.3, indicating a slightly horizontal shape. HOR_{MID} has five outliers. For both horizontality measures, only a few families show values higher than 2.



(a) HOR_{MAX} **(b)** HOR_{MID}

Figure 6.23: Boxplots and histograms of the horizontality measures.

6.3.4.4 Compactness Measures

Several indices are calculated to measure the geometric compactness of the language families. For vector data, a circle is the most compact shape (Montero and Bribiesca, 2009; Li et al., 2014). Therefore, the following measures are also called circularity measures. There are different measures to assess the compactness of shapes. One category consists of area-perimeter measures, which are rather simple measures based on the perimeter and the area of a polygon. Another category consists of reference shape measures comparing a polygon shape to the most compact shape encompassing the polygon.

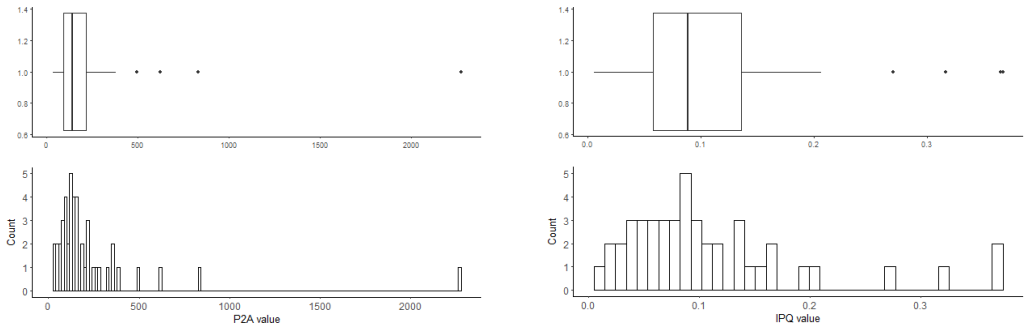
Area-Perimeter Measures

In this subsection the classical ratio of the squared perimeter and area (referred to as P2A) and the Isoperimetric Quotient (IPQ) are calculated. There are other area-perimeter measures, but most of them are closely related. For instance, sometimes the reciprocal of the IPQ is used, which is referred to as normalized P2A. The square root of the normalized P2A is also used frequently and is called corrected P2A (CPA).

P2A The P2A measure is also called shape factor or roundness measure and it is the most widely used compactness measure (Montero and Bribiesca, 2009). It compares the squared perimeter of a shape to the area of this same polygon (Montero and Bribiesca, 2009): $P2A = perimeter^2/area$. A circle has a P2A value of 12.6. The less compact the shape, the higher the value.

IPQ The IPQ represents the ratio of the polygon area to the area of a circle with the same perimeter. Cox (1927) used this measure to assess the roundness of sand grains, but nowadays this measure is widely applied in different contexts. Compactness is assessed "by the degree to which the ratio of the area to the circumference approaches the same ratio for a circle" (Cox, 1927, p. 180). In case of a circle, $area/perimeter^2$ equals $1/(4 \cdot pi)$. Multiplying this equation with $4 \cdot pi$ results in $(4 \cdot pi \cdot area)/perimeter^2 = 1$ and the measure to calculate is thus: $IPQ = (4 \cdot pi \cdot area)/perimeter^2$. The resulting values range from 0 to 1, whereas 1 stands for perfect roundness.

The resulting P2A values range from 34 to 2'270 and most values are found between 40 and 250 (see figure 6.24a). Four outliers show values higher than 500, indicating extreme non-compactness. Figure 6.24b shows the IPQ values ranging from 0.006 to 0.367. IPQ values higher than 0.2 are rare and most of the families have values between 0.04 and 0.1 indicating non-compact shapes. There are four outliers indicating compactness. Due to the way the measures are calculated, the values show a negative exponential relationship. Thus, regarding the ranking of families, they show the same order, with Austronesian being the least compact and Pano-Tacanan being the most compact shape.



(a) P2A **(b)** IPQ

Figure 6.24: Boxplots and histograms of the area-perimeter measures.

Reference Shape Measures

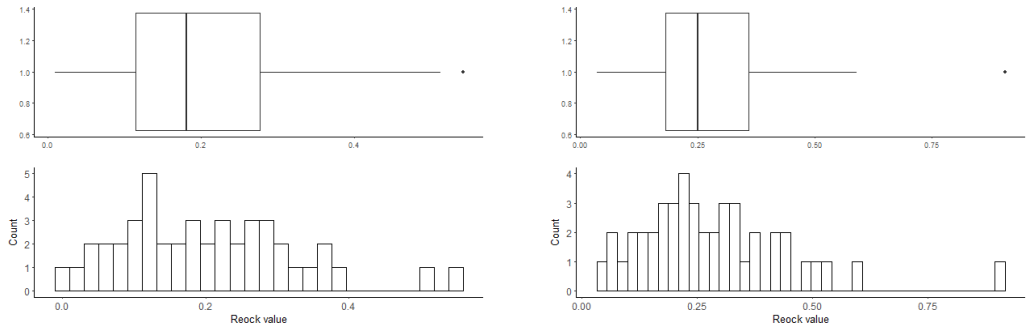
The measures in this class compare the polygon under investigation to its respective minimum standard reference shape, a circle in the case of roundness, that encompasses the polygon (Li et al., 2014). In this thesis, the Reock measure is calculated and it is additionally corrected for the landmass.

Reock Measure (REOCK) The Reock value of a polygon is calculated by dividing the actual shape area by the area of the smallest circle encompassing the shape, i.e. the circumcircle (Reock, 1961): $REOCK = area_{\text{polygon}}/area_{\text{circumcirle}}$. The resulting values range from zero to one, whereas one indicates a circular shape. In figure 6.25 the language family polygon of Benue-Congo is shown in orange and its circumcircle in blue. The shape area (9'400'500 km^2) divided by the area of the circumcircle (26'005'700 km^2) results in a Reock value of 0.361.

Corrected Reock Measure (REOCK$_{CORR}$) The Reock measure does not account for underlying geographical characteristics like landmass or ocean. Some circumscribing circles thus encompass a relatively large part covered by water. This is corrected by replacing the denominator by the area of the landmass lying within the smallest circumscribing circle: $REOCK_{\text{CORR}} = area_{\text{polygon}}/area_{\text{circumcirle.landmass}}$. Benue-Congo has a REOCK$_{\text{CORR}}$ value of 0.589 due to the reduced area of the circumcircle (26'006'000 km^2 (blue and orange area in figure 6.25) reduced to 15'958'000 km^2), indicating a more circular shape than when using the REOCK (0.361).

Figure 6.25: Illustration of the calculation of the Reock meausures for the language family Benue-Congo. The language family area is depicted in orange and the blue circle illustrates the circumcircle. Combining the blue and orange area results in the area within the circumcirle lying on the landmass which is used for the calculation of $\text{REOCK}_{\text{CORR}}$. *Data: Natural Earth.*.

The values of the Reock measure range from 0.009 to 0.542 (see figure 6.26a). The $\text{REOCK}_{\text{CORR}}$ values show a different picture with values ranging from 0.035 to 0.906 (see figure 6.26b). Due to the subtraction of the water area lying in the circumcircle, the values show more compact results in general. Both distributions result in normally distributed data with mean values of 0.20 and 0.28 and standard deviations of 0.12 and 0.16 respectively.



(a) REOCK **(b)** $\text{REOCK}_{\text{CORR}}$

Figure 6.26: Boxplots and histograms of the reference shape measures.

6.3.4.5 Correlation Matrix

The different geometric factors of the language family areas are not independent of each other. Table 6.7 shows the correlations among the different factors. The strongest correlations exist between the two horizontality measures and the two Reock measures. This is not surprising because the basis of their calculation is the same. Further, the areal size and the perimeter correlate quite strongly and the IPQ depends on the area and the perimeter (out of which it is calculated). It is rather surprising that there is no strong correlation between the cardinal size and geometric properties.

Table 6.7: Correlation matrix of the geometric measures. The correlation coefficient (Spearman's *rho*) between each pair is depicted.

Geometric Properties	AREA	PERI	HOR_{MAX}	HOR_{MID}	IPQ	REOCK	$REOCK_{CORR}$
SIZE	0.33	0.42	0.12	0.13	-0.48	-0.09	-0.05
AREA		0.95	0.15	-0.05	-0.52	-0.06	-0.15
PERI			0.21	0.03	-0.74	-0.21	-0.28
HOR_{MAX}				0.94	-0.24	-0.50	-0.42
HOR_{MID}					-0.19	-0.50	-0.41
IPQ						0.51	0.51
REOCK							0.82

This chapter discussed the calculation of the 28 geographical factors for each language family area. These encompass the cardinal size, environmental characteristics, neighbourhood measures and geometric properties. The following section links these factors to the transition rates by means of a correlation analysis.

Chapter 7

Results

This chapter presents the results of the correlation analyses between language change and geographical factors. Two kinds of analyses are performed, namely an analysis on an aggregated level and on a feature level. The former investigates phonological and grammatical change on the level of language families. This means that an average transition rate for phonological and grammatical features is calculated for each family. These rates are then compared to the geographical factors. It would also be possible to investigate the standard deviation of the rates. This is not done because it would mean that a different concept in language evolution theory is examined, namely by which geographical setting variability in evolution is promoted. The correlation analysis on the feature level compares the transition rates of each linguistic feature with each geographical factor.

On both levels, a linear correlation analysis is performed. As several geographical factors do not follow a normal distribution, the rank correlation by Spearman is performed. This is done for all factors including the ones that follow a normal distribution in order to be able to compare the results. As several hypotheses are tested, the statistical significance resulting from the rank correlation is not informative because various factors interact (see e.g. Bonferroni correction). Therefore, in this analysis, significance values are not considered, but the correlations between geographical factors and language change are qualitatively assessed. This is appropriate regarding that the next step (which is not performed in this thesis) would be the statistical modelling of language contact. For both levels, relationships with an absolute correlation coefficient of at least 0.2 are shown. In the context of contact-induced language change this is a suitable value because there are a lot of social, economic and political factors influencing language contact.

As mentioned in section 5.2, language families are of different cardinal size, which makes them hard to compare. Therefore, the analyses are not only performed for datasets including all families but also for datasets without the four biggest families. The latter are from now on referred to as WB families. The four outliers regarding cardinal size (see figure 6.7) are defined as big families: Austronesian (1137 languages), Benue-Congo (945 languages), Sino-Tibetan (448 languages) and Indo-European (405 languages). The exclusion of these families yields results that are less affected by large cultural and historical processes.

7.1 Correlations on Aggregated Level

This section presents the results of the correlation analysis between the average transition rates of grammatical and phonological features and geographical factors. As mentioned before, the distinction between phonology and grammar is made because it is assumed that the resulting correlations may differ.

To see whether the average values of the transition rates of the phonological and grammatical features represent the data of a family well, they are tested for normal distribution using the Kolmogorov-Smirnov test. In order to guarantee a certain representativeness of language change, families with less than five features are not incorporated into the analysis. Additionally, five observations are the prerequisite for performing the Kolmogorov-Smirnov test. Including families with less data would thus allow them to influence the results of the analysis with the dataset that also includes non-normally distributed data. Data including non-normally distributed transition rates is tested as well because of the small number of language families with normally distributed transition rates for grammatical features. This small number may influence the results of Spearman's rank correlation.

The correlation analysis is performed on eight different datasets: the main division of the data is the separation of phonology and grammar, the second division is based on the normal distribution of the transition rates within a family (all data vs only normally distributed data) and within this division, it is differentiated between WB families and all families. The results of the correlation analysis are presented in table 7.1. In the following, this table is discussed by first outlining the number of families incorporated into the analysis and the number of correlations. Second, the geographical factors which correlate are described. Third, the correlations between these factors and phonological and grammatical change are elaborated on and fourth, the different datasets are compared.

7.1.1 Datasets and Number of Correlations

There is a remarkable difference in the data for the phonological and grammatical features. Phonological features have been collected more widely, which is reflected in the large number of families containing at least five features. In total, 36 (out of 45) families contain rates of at least five features. Four of them are the biggest families resulting in 32 WB families with enough data. Of the 36 families, 25 have normally distributed rates, of which 23 are WB families (thus, only two of the four biggest families have normally distributed data). Most of the grammatical features have only been collected for large language families, while a lot of the smaller families only have values for a few features. In total, 20 families have values for at least five grammatical features. These 20 families include the four biggest families and 16 WB families. Of the 20 families, 14 have normally distributed data. 13 of these normally distributed families are WB families, i.e. only one of the four biggest families has normally distributed transition rates. Based on these numbers, the results of the correlation analysis of the phonological features can be trusted more than the results of the analysis of

Table 7.1: Results of the correlation analysis between the average values of the phonological and grammatical transition rates per family and the geographical factors. Correlations with an absolute *rho* of at least 0.2 are shown. Values above 0.3 are marked in bold.

		Phonology				**Grammar**			
	only normally distributed	yes		no		yes		no	
	families	all	WB	all	WB	all	WB	all	WB
	n	25	23	36	32	14	13	20	16
Lingusitic factor	SIZE					0.24	0.26		0.24
Environmental characteristics	$PREC_{AV}$	**0.48**	**0.48**	0.28	0.27				
	$PREC_{SD}$					**0.44**	**0.40**	**0.33**	**0.39**
	$TEMP_{MIN.AV}$	**0.51**	**0.49**	**0.30**	0.28				
	$TEMP_{MIN.SD}$					0.22	0.24		0.27
	$TEMP_{MAX.SD}$	**-0.30**	-0.23						
	ALT_{AV}	**-0.33**	**-0.32**			**0.36**	**0.39**	0.24	**0.41**
	ALT_{SD}	**-0.31**	-0.25	-0.23	-0.20	0.28	**0.31**		**0.32**
	TRI_{AV}	-0.23	-0.22	-0.24	-0.23				
	TRI_{SD}	-0.22							
Neighbourhood measures	$ADJ_{LANG.TOT}$					0.25	0.29		0.29
	$ADJ_{LANG.SF}$	**-0.38**	**-0.40**	**-0.39**	**-0.42**				
	$ADJ_{LANG.DF}$	**0.44**	**0.43**	**0.48**	**0.48**				
	ADJ_{FAM}					**0.54**	**0.56**	**0.49**	**0.60**
	$PD_{100.TOT}$			-0.23	-0.28				
	$PD_{100.SF}$	-0.21	-0.22	-0.28	**-0.31**				
	$PD_{500.TOT}$			-0.22	-0.27				
	$PD_{500.SF}$	-0.25	-0.25	**-0.30**	**-0.34**				
Geometric properties	AREA				0.23				
	PERI				0.26				
	HOR_{MAX}								-0.22
	IPQ	-0.28	**-0.40**	-0.26	**-0.35**				
	REOCK		-0.21						
	$REOCK_{CORR}$	**-0.32**	**-0.38**	-0.28	**-0.33**				

the grammatical features due to the small number of families with at least five grammatical features.

Across the eight datasets, there are 77 out of 224 (28 factors*8 datasets) possible correlations in total, of which 37 have an absolute *rho* of at least 0.3 (marked in bold in table 7.1). The correlations go up to 0.60, which is a high value considering that the datasets consist of average values. The average phonological transition rates show more correlations with the geographical factors than the average grammatical transition rates. Both phonological datasets with normally distributed data have 13 correlations. The datasets also incorporating non-normally distributed data show 12 (all families) and 14 (WB families) correlations. The grammatical datasets with normally distributed data result in seven correlations whereas the other datasets have three (all families) and eight (WB families) correlations.

Not only the number of correlations, but also the strength of the correlations differs across the datasets. The phonological datasets generally show stronger correlations for the normally distributed data, which supports the respective correlations, because correlations with normally distributed data can be trusted more because they represent the data of a family better than the datasets that include non-normally distributed data as well. The grammatical datasets, however, show a mixed picture. Moreover, there is a general difference in the resulting correlations if all families are included or only the WB families. The picture is the same within both the normally distributed datasets and the datasets also incorporating families with non-normally distributed transition rates. For the phonological datasets, the environmental factors show weaker correlations for the WB families. For the neighbourhood and geometric factors, the WB datasets show stronger correlations apart from one correlation ($ADJ_{LANG.DF}$). For the grammatical datasets, the correlations of the WB families are always stronger with the exception of the normally distributed dataset, in which all families correlate more strongly with $PREC_{SD}$. As the normally distributed datasets differ only in one family, the results are similar.

7.1.2 Geographical Factors and Correlations

Four geographical factors ($TEMP_{MAX.AV}$, $PD_{100.DF}$, $PD_{500.DF}$, and HOR_{MID}) do not correlate at all. TRI_{SD}, HOR_{MAX}, REOCK, AREA, and PERI correlate with one dataset only. $PD_{100.TOT}$ and $PD_{500.TOT}$ each correlate with two datasets. The majority of geographical factors correlate with three or four datasets. ALT_{AV} correlates with six and ALT_{SD} with seven datasets. The focus of the following elaboration lies on the factors that correlate at least two or three times. Moreover, the correlations between the geographical factors and the transition rate datasets are stable within the subdivision of phonology and grammar. That is, if a geographical factor correlates with more than one dataset, while the strength of the relationships may differ, their orientation, positive or negative, remains the same. How the single geographical factors correlate with phonological and grammatical change is addressed in the next section.

7.1.3 Phonological and Grammatical Change

Phonological change shows several correlations with geographical factors. Regarding environmental factors, phonological change shows positive correlations with the average measures of the climatic factors $PREC_{AV}$ and $TEMP_{MIN.AV}$. Both factors show relatively high correlations and correlate with all datasets. The factor $TEMP_{MAX.SD}$ correlates negatively with the normally distributed datasets. Further, all topographical measures show negative correlations: ALT_{SD} and TRI_{AV} correlate with all four datasets, but the correlations are rather weak. ALT_{AV} shows strong correlations with the normally distributed datasets, while TRI_{SD} shows only one weak correlation. In respect of the neighbourhood measures, phonological change correlates positively with the increasing number of neighbours from different families ($ASJ_{LANG.DF}$) and negatively with the increasing number of neighbours of the same family ($ADJ_{LANG.SF}$, $PD_{100.SF}$, $PD_{500.SF}$). Both adjacency measures correlate highly for all four datasets. PD_{SF} measures show weaker correlations, but still for all datasets. Further, $PD_{100.TOT}$ and $PD_{500.TOT}$ show rather weak negative correlations with the datasets that incorporate also non-normally distributed average rates. Regarding geometric properties, there are weak correlations with AREA and PERI for one dataset and rather strong negative correlations for the compactness measures. IPQ and $REOCK_{CORR}$ show relatively high correlations for all datasets, while REOCK only correlates weakly with the dataset incorporating normally distributed WB families.

As mentioned before, grammatical change shows fewer correlations with geographical factors than phonological change: cardinal size correlates with three datasets. There is no correlation for the dataset for all families incorporating non-normally distributed data. There are correlations between grammatical change and environmental factors. The correlation with $PREC_{SD}$ is positive and quite strong for all four datasets. The only temperature factor correlating is $TEMP_{MIN.SD}$: the correlation is rather weak but exists for three datasets. The topographical measure ALT_{AV} shows quite strong correlations for all datasets. ALT_{SD} shows positive correlations for three datasets. Grammatical change further correlates positively with the neighbourhood measures ADJ_{FAM} and $ADJ_{LANG.TOT}$. For the former, the correlation is strong and observable for all four datasets and for the latter it is rather weak and observable for three datasets. Regarding geometric properties of language family areas, only HOR_{MAX} correlates negatively with the WB families of the dataset that also incorporates non-normally distributed data.

In this section, the results of the correlation analysis on the aggregated level were described. In the following, the results of the correlation analysis on the feature level are outlined.

7.2 Correlations on Feature Level

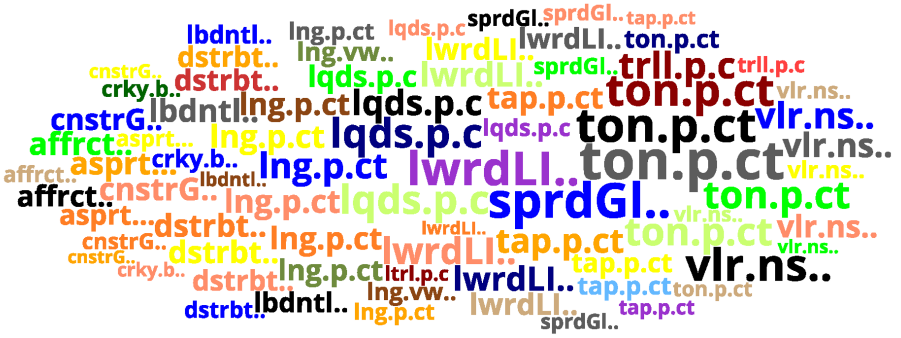
This section presents the results of the feature level correlation analysis, during which each linguistic feature is compared to each geographical factor. The correlation analysis is performed for features with rate estimates for at least 15 language families. Although this minimum value is still low for the conduction of a correlation analysis, it guarantees a certain stability of the results. The analysis is performed for two different datasets, namely one taking into account all families and the other incorporating only the WB families.

To facilitate the legibility of the results, geographical factors belonging to the same of the three classes (environmental characteristics, neighbourhood measures, and geometric properties) are combined into factor classes provided that they show high correlations between each other (for correlation matrices of geographical factors see tables 6.3, 6.6, and 6.7). Within these factor classes, the highest correlation is selected. For instance, if a feature shows a correlation of -0.3 with IPQ and of -0.2 with REOCK, the resulting class COMPACTNESS has a correlation value of -0.3. The classification is shown in table 7.2. This table also serves as a legend for the word clouds depicting the results of the correlation analysis (see figures 7.1 and 7.2). The former figure depicts the result of the analysis incorporating the dataset including all families, while the latter depicts the results of the analysis incorporating only the WB families. For both datasets, the results are structured by phonological and grammatical features and by positive and negative correlations. This results in four word clouds per dataset. The words depict the name of the respective feature or its abbreviation, respectively (for full names see table A.1). The colour of the words represents the factor class it correlates with and the size of the words reflects the relative strength of the relationship between the respective feature and the factor class within the world cloud. This means that the sizes of the words cannot be compared across the four figures, but only within the respective figure.

Table 7.2: Classification of the factors into factor classes. This table also serves as a legend for figures 7.1 and 7.2.

Factor Class	Factors
SIZE	SIZE
$PREC_{AV}$	$PREC_{AV}$
$PREC_{SD}$	$PREC_{SD}$
$TEMP_{MAX.AV}$	$TEMP_{MAX.AV}$
$TEMP_{MAX.SD}$	$TEMP_{MAX.SD}$
$TEMP_{MIN.AV}$	$TEMP_{MIN.AV}$
$TEMP_{MIN.SD}$	$TEMP_{MIN.SD}$
TOPO	ALT_{AV} ALT_{SD} TRI_{AV} TRI_{SD}
$ADJ_{LANG.TOT}$	$ADJ_{LANG.TOT}$
$ADJ_{LANG.SF}$	$ADJ_{LANG.SF}$
$ADJ_{LANG.DF}$	$ADJ_{LANG.DF}$
$ADJ_{LANG.FAM}$	ADJ_{FAM}
PD_{TOT}	$PD_{100.TOT}$ $PD_{500.TOT}$
PD_{SF}	$PD_{100.SF}$ $PD_{500.SF}$
PD_{DF}	$PD_{100.DF}$ $PD_{500.DF}$
AREA/PERI	AREA PERI
HOR	HOR_{MAX} HOR_{MID}
COMPACTNESS	IPQ REOCK $REOCK_{CORR}$

(a) Phonological features that correlate positively.



(b) Phonological features that correlate negatively.

(c) Grammatical features that correlate positively.

(d) Grammatical features that correlate negatively.

Figure 7.1: Correlations of linguistic features with geographical factors. All language families are incorporated. Correlations with an absolute *rho* above 0.2 are shown. The colour represents the factor class (table 7.2) and the size reflects the relative strength of the correlation. *Made with wordle (http://www.wordle.net/).*

(a) Phonological features that correlate positively.

(b) Phonological features that correlate negatively.

(c) Grammatical features that correlate positively.

(d) Grammatical features that correlate negatively.

Figure 7.2: Correlations of linguistic features with geographical factors. WB families are incorporated. Correlations with an absolute *rho* above 0.2 are shown. The colour represents the factor class (table 7.2) and the size reflects the relative strength of the correlation. *Made with wordle (http://www.wordle.net/).*

7.2.1 General Description

There is a big difference in the number of phonological and grammatical features because most of the grammatical features are only collected for a few large families. Hence, there are only a few grammatical features with rates for at least 15 language families. The dataset incorporating all families contains six grammatical features, one of which has 19 rates and the other five have 15 or 16 rates. The same dataset contains 16 phonological features, of which 15 have more than 20 rates. The omission of the large families leads to the exclusion of five grammatical features and of one phonological feature.

Regarding the phonological features, there are 128 out of 288 (16 features*18 factor classes) possible correlations for the dataset including all families (see figures 7.1a and 7.1b). Incorporating only the WB families results in 142 out of 270 (15 features*18 factor classes) possible correlations (see figures 7.2a and 7.2b). Regarding the grammatical features, incorporating all families results in 59 of 108 (6 features * 18 factor classes) possible correlations (see figures 7.1c and 7.1d). If only the WB families are incorporated, 12 correlations can be observed for the one remaining grammatical feature (see figures 7.2c and 7.2d).

For both datasets, every feature that is incorporated shows at least three correlations and all factor classes correlate as well. Some factor classes, however, correlate with more features than others. For instance, $PREC_{SD}$, TOPO, and COMPACTNESS show several correlations for both phonological and grammatical features. A more detailed analysis of which factors or factor classes respectively influence phonological and grammatical features is provided in the following section.

7.2.2 Phonological and Grammatical Change

For the phonological features, the resulting trends of the correlation analysis regarding negative and positive correlations are the same for both datasets. Except for the correlations with $TEMP_{MAX.SD}$ and HOR, omitting the four largest families results in correlations that enhance the pattern indicated by the correlations in the data incorporating all families. In both datasets, some features show more correlations in total, but they are not considerably stronger than the correlations of other features. All factor classes correlate for both datasets, however, only for a few factor classes a pattern is discernible regarding positive or negative relationships between the change of phonological features and geography. Regarding the cardinal size no correlation pattern is discernible. $TEMP_{MAX.AV}$ and $TEMP_{MAX.SD}$ are the only climatic measures that show clear patterns; the former shows a positive and the latter shows a negative relationship. $TEMP_{MIN.AV}$ further indicates a slightly positive relationship. Regarding TOPO, the pattern shows a trend towards a negative relationship, although there are some positive correlations as well. Regarding the neighbourhood measures, there are several positive and negative correlations. For ADJ_{FAM} no clear pattern is discernible. ADJ_{TOT} shows no clear picture regarding its correlations, while ADJ_{SF} correlates mainly

negatively and ADJ_{DF} shows strong positive correlations. All PD classes show a pattern indicating negative relationships. Further, AREA/PERI shows no pattern for the dataset incorporating all families, but it correlates positively for the WB families. For HOR, the pattern indicates a slightly positive relationship. COMPACTNESS shows relatively strong negative correlations.

The correlations between the factor classes and grammatical features of the WB family dataset are not investigated further as only one feature was incorporated into the correlation analysis. Hence, only the results of the analysis with all families are investigated in more detail. The different grammatical features show roughly the same number of correlations. Due to the small number of grammatical features, clear patterns are not discernible. For the environmental factors, there is a tendency towards a positive relationship with the TOPO class. There are weak positive relationships with $ADJ_{LANG.DF}$ and PD_{DF}. Some tendencies towards a negative relationship exist with AREA/PERI and horizontality.

This chapter described the results of the correlation analyses on both the aggregated and the feature level. The next chapter addresses the interpretation of these outcomes.

Chapter 8

Interpretation

This chapter interprets the results of the correlation analyses, i.e. the meaning of the correlations between phonological and grammatical change and the geographical factors. Subsequently, the appropriateness of the computed geographical factors is reflected on.

For the interpretation of the results of the feature level correlation analysis experts of linguistics were consulted. The discussion revealed that the patterns resulting on this level are hard to describe as no signal is detectable. The noise on this level may be due to uncertainties that added up: first, the phylogenetic tree used is the best for the respective family, how accurate this tree actually is, however, cannot be determined, because true phylogenies are not available (Nichols and Warnow, 2008). Second, by the estimation of the transition rates further uncertainties are introduced. Hence, an in-depth analysis of how single features correlate with geographical factors is not provided. The results of this level are only consulted to confirm trends regarding correlations of phonological and grammatical change that have been found on the aggregated level.

The correlations that are observable on both analysis levels are consistent in a way that they follow the same pattern, i.e. they are either positive or negative. This indicates that the geographical factors show a consistent pattern.

8.1 Linguistic Interpretation

This section interprets the correlations between geographical factors and language change presented in chapter 7 by linking them to the expectations outlined in section 4.3. As described in chapter 7, the four datasets produce different results, whereby the correlations with the normally-distributed datasets can be trusted more than the correlations with the datasets also incorporating non-normally distributed data. In general, however, if a correlation between phonological or grammatical change and a geographical factor is high, it exists for at least three datasets. This confirms the existence of the relationship. For the interpretation, the focus lies on these correlations. Interpreting the correlations, it has to be kept in mind that the correlations for grammatical change are more uncertain because they are based on less data.

8.1.1 Cardinal Size

Cardinal size only correlates with grammatical change. The correlation on the aggregated level suggests a positive relationship. This corresponds to the expectation that the more languages a family comprises, the higher is the rate of change. As this factor is not geographical *per se* and does not operationalise the conceptual factors important for contact-induced language change, it will not be discussed further. In a further statistical modelling, however, it might reveal informative results when combined with other factors.

8.1.2 Environmental Characteristics

The environmental characteristics operationalising the probability of contact show different pictures regarding their influence on language change. Climatic average measures only correlate with phonology: both analysis levels suggest a positive relationship between phonological change and the climate average measures $PREC_{AV}$ and $TEMP_{MIN.AV}$. This indicates that the more precipitation and the higher the average minimum mean temperature in a language family area, the more probable is contact-induced language change. This contradicts Nettle's (1998) ecological risk theory stating that these circumstances promote self-supply, which leads to less contact. The causal relationship between temperature and latitude influencing this correlation is addressed in section 8.2.1.

Climate variability shows two positive relationships with grammar, namely for $PREC_{SD}$ and $TEMP_{MIN.SD}$. This indicates that the higher climate variability within a language family area is, the more contact-induced language change occurs. This is in accordance with the expectation that due to altered climatic conditions during the migration of a group, contact with other groups is necessary and the chance of being in contact with other groups increases (Güldemann, 2010; Diamond, 1997). There is a negative relationship between phonological change and $TEMP_{MAX.SD}$, contradicting the findings for grammatical change. Critical elaborations on the $TEMP_{MAX}$ measures are discussed in section 8.2.2.2.

Regarding topographical complexity, the relationships with phonological and grammatical change contradict: the correlation with phonology is negative, while the correlation with grammar is positive. The relationships are strong for both structural domains. The pattern of the feature level analysis supports these correlations. Regarding phonology, this relationship is observable for all four topographical measures. In comparison, regarding grammatical change, the positive relationship is suggested only by correlations with ALT_{AV} and ALT_{SD}, which are conceptually more critical measures (see further explanation in section 8.2.2.2). These results suggest that isolation, which is favoured by a rough terrain, leads to a slow change in phonology but to a fast change in grammar. The change of phonology corresponds to the expectations of the topography theory (Stepp, Castaneda, and Cervone, 2005), while fast grammatical change does not support the theory.

8.1.3 Neighbourhood Measures

The neighbourhood measures mapping the potential of language contact show distinct patterns: for phonological change, all adjacency measures show a consistent pattern for both levels: change is not dependent on ADJ_{FAM} and $ADJ_{LANG.TOT}$, but there is a high negative correlation with $ADJ_{LANG.SF}$ and a high positive correlation with $ADJ_{LANG.DF}$. This means that the more genetically unrelated neighbours there are, the faster language change occurs. If a lot of neighbours belong to the same family, however, phonological language change is slow. This corresponds to the expectation that structural similarity between languages, which is generally high for related languages, may minimise contact-induced language change. Regarding grammatical change, the results of the aggregated level indicate a strong positive correlation between change and both $ADJ_{LANG.FAM}$ and $ADJ_{LANG.TOT}$. This is a different pattern than for phonology. Nevertheless, it also corresponds to the expectations that in general, the potential of contact-induced language change is higher if the number of neighbours is high. For a rapid change of phonological features it seems to be important that adjacent languages belong to a different language family, while for grammatical change it is only important that there are a lot of neighbours.

Regarding the PD measures, there is a negative correlation between phonological change and the PD_{SF} measures, indicating that the higher the density of languages from the same family, the less language change occurs. Further, the results of both levels indicate a negative relationship with the PD_{TOT} measures. The relationship with PD_{SF} confirms the findings of the $ADJ_{LANG.SF}$ measure. The correlation with PD_{TOT}, however, contradicts the expectation that the more neighbours, the higher the probability of language contact. This, however, may be due to the concept these measures are based on, which is discussed in section 8.2.2.3.

8.1.4 Geometric Properties

Several geometric properties (reflecting the probability of language change being effective) correlate with language change. For phonology, both analysis levels show a slight tendency towards a positive relationship with AREA and PERI. This indicates that the bigger the areal size and thus also the perimeter of a language family (*rho* = 0.94 between AREA and PERI), the more probable is language contact. This was expected (Currie and Mace, 2009), but the relationship is only indicated and thus quite uncertain.

Horizontality does not have an influence on phonological change. There is only a slight indication of a negative relationship with grammatical change, which is very uncertain. Shape compactness, however, shows a clear pattern on both levels for phonological change: The relationship is negative, indicating that the more compact the area, the less change occurs. This was expected as shape compactness of a language family is associated with stable social structures, which in turn suggests that in case language contact occurs, it is less effective (Diamond, 1997; Trudgill, 2010).

8.2 Reflection on the Geographical Factors

This section addresses the informative value of the geographical factors. In advance, however, the correlation of geographical factors with latitude is addressed.

8.2.1 Latitude as a Determinant

Given the latitudinal gradient in language and language family diversity, several of the geographical factors correlate with latitude. For this examination, the latitude of a family is defined as the latitude of the midpoint of the maximum and minimum latitude of a language family polygon. Spearman's rank correlation (see table 8.1) reveals that several geographical factors show correlations with latitude with an absolute *rho* higher than 0.5: there are positive correlations for temperature variability measures, namely $TEMP_{MAX.SD}$ and $TEMP_{MIN.SD}$. Further, language family areas are smaller in equatorial regions and several neighbourhood measures (in particular PD measures) correlate negatively with latitude confirming the high language density at lower latitudes. All those relationships, however, are not causal. For instance, the correlations of the temperature variability measures can be explained by smaller language family areas, the size of which is dependent on language and language family diversity, which is high in the equatorial region.

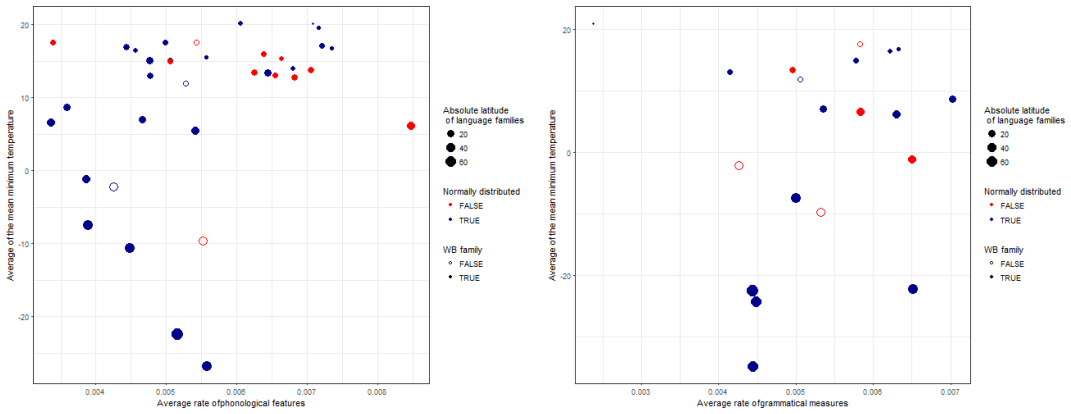
The strong negative relationship between $TEMP_{MIN.AV}$ and latitude (*rho* = -0.89), however, is causal. There is no strong correlation between $TEMP_{MAX.AV}$ and latitude. This can be explained by high summer temperatures of continental climates in higher latitudes. $PREC_{AV}$ also correlates strongly, but the influence of wind systems etc. is too high to assume a causal relationship. Therefore, in the following, the relationship between $TEMP_{MIN.AV}$ and language change is investigated by taking into consideration latitude as a determinant. For this, only the aggregated level is looked at, because in-depth investigations on the feature level are not performed due to the reasons mentioned above (see section 8.1).

Table 8.1: Correlation of the geographical factors with latitude.

Geographical factor	***rho***
SIZE	-0.18
$PREC_{AV}$	-0.76
$PREC_{SD}$	-0.35
$TEMP_{MIN.AV}$	-0.89
$TEMP_{MIN.SD}$	0.57
$TEMP_{MAX.AV}$	-0.26
$TEMP_{MAX.SD}$	0.65
ALT_{AV}	0.26
ALT_{SD}	0.28
TRI_{AV}	0.10
TRI_{SD}	0.02
$ADJ_{LANG.TOT}$	-0.57
$ADJ_{LANG.SF}$	-0.28
$ADJ_{LANG.DF}$	-0.28
ADJ_{FAM}	-0.17
$PD_{100.TOT}$	-0.55
$PD_{100.SF}$	-0.49
$PD_{100.DF}$	-0.63
$PD_{500.TOT}$	-0.6
$PD_{500.SF}$	-0.4
$PD_{500.DF}$	-0.66
AREA	0.50
PERI	0.40
HOR_{MAX}	0.25
HOR_{MID}	0.04
IPQ	0.00
REOCK	0.11
$REOCK_{CORR}$	0.14

Temperature and Latitude

Due to the causal relationship between temperature and latitude it is expected that the positive relationship between transition rates and $TEMP_{MIN.AV}$ is caused by outliers, i.e. by high latitude families showing a low minimum temperature. It is assumed that not incorporating these outliers would result in a rather negative relationship. To investigate this assumption, $TEMP_{MIN.AV}$ values are plotted against the transition rates of phonological and grammatical features respectively, whereby the language families are displayed as points (see figure 8.1). The size of the points depicts the absolute latitude, their colour designates if the transition rates of this family follow a normal or non-normal distribution and the shape indicates whether or not it is a WB family. Hence, for both grammar and phonology, all four datasets used for the correlation analysis are depicted in these figures.



(a) Phonology **(b)** Grammar

Figure 8.1: Crossplot of TEMP.MIN.AV and the average rate of change of phonological (a) and grammatical (b) features. The language families are depicted as points. The point size represents the absolute latitude: The colours show whether the rates are normally or non-normally distributed and the filled points represent the WB families.

When investigating the influence of the latitude on the correlation between $MIN_{TEMP.AV}$ and phonological change (see figure 8.1a), it becomes obvious that the high latitude families have an influence on the strong positive relationship (large points in the lower left of the plot). The low latitude families (0-25°) cluster in the upper part of the plot showing high $MIN_{MEAN.AV}$ values. When incorporating both normally and non-normally distributed data (red and blue points), there is no relationship without the high latitude families. The relationship only becomes positive when families with latitudes above 30° are included. When only the normally-distributed data is taken into account (blue points), the relationship stays positive also without the high latitude families. The relationship gets stronger when the latitude of the included language families increases. In both cases the WB families do not considerably influence the results.

When investigating $MIN_{TEMP.AV}$ and grammar (see figure 8.1b), high latitude families show low $MIN_{TEMP.AV}$ values. The relationship, however, is weak and negative for all four

datasets. Disregarding the high latitude families results in a stronger negative relationship for all datasets. For the two datasets that include non-normally distributed data, the negative relationship is strongest when families up to 30° are taken into account. For the two datasets including only normally distributed data, the relation is strongest for families with latitudes below 20°. It has to be considered that these values are based on only six or seven language families making it critical to draw a conclusion.

These elaborations suggest that the causal relationship between temperature and latitude is important and that it shifts the relationship between temperature and transition rates into the direction of a positive relationship: regarding the relationship of temperature and phonological change this means that the relationship shifts from a weak positive to a positive correlation. Regarding grammatical change, the relationship shifts from a negative to a weak negative correlation. The influence, however, is not as strong as expected.

8.2.2 Suitability of the Geographical Factors

This section aims at assessing the appropriateness of the geographical factors by investigating the correlations among the geographical factors. In doing so, relationships with an absolute *rho* of at least 0.5 are regarded as correlations. Furthermore, the results of the correlation analyses with language change are taken into account.

8.2.2.1 Correlations of Geographical Factors Between Classes

Most of the geographical factors correlate only with other factors of the same factor class (environmental characteristics, neighbourhood measures, geometric properties). The complete correlation matrix can be viewed in the appendix (table A.2). Cardinal size correlates only with neighbourhood measures. This indicates that the geographical factors are suitable operationalisations of the conceptual factors. There are some exceptions for AREA and PERI: within the geometric measures, they only correlate with the IPQ (which puts AREA and PERI into a relation, see section 6.3.4.4). They show negative correlations with the average climatic measures ($PREC_{AV}$, $TEMP_{MIN.AV}$) and positive correlations with temperature variability measures ($TEMP_{MIN.SD}$, $TEMP_{MAX.SD}$). As indicated before, this can be explained by the correlation of AREA and latitude. A further exception is the positive correlation between $PREC_{AV}$ and several PD neighbourhood measures. These relationships suggest that the more precipitation occurs in a language family area, the higher is the language density. This relation has been found before and can be connected to the theory of Nettle (1998), which suggests that high amounts of precipitation lead to a high language diversity as precipitation is important for food production and thus for the self-supply of groups. This is further supported by an indication of a relationship between temperature and language density (correlation between $PD_{500.DF}$ and $TEMP_{MIN.AV}$).

In the following, the different geographical factors are discussed in more detail regarding correlations with other geographical factors and also based on the resulting correlations interpreted above.

8.2.2.2 Environmental Characteristics

Investigating the correlations among the environmental factors (see table 6.3), they match the expectations and they can be explained by the areal size of language families correlating with latitude as described above. One remarkable result is that $TEMP_{MIN.AV}$ and $TEMP_{MAX.AV}$ do not correlate and that $TEMP_{MAX.AV}$ does not correlate with latitude either. Regarding Nettle's theory, $TEMP_{MIN.AV}$ plays a more important role than $TEMP_{MAX.AV}$ because it is rather the minimum than the maximum temperature that hampers food production. $PREC_{AV}$ and $TEMP_{MIN.AV}$ correlate strongly (*rho* = 0.84). Their correlation with latitude confirms that they measure what they are intended to. A connection between precipitation and temperature in terms of statistical modelling, however, might lead to different results regarding language contact because this connection would take the existence of hot, dry areas such as deserts into consideration.

As the maximum mean temperature seems to represent a different concept than the minimum mean temperature, $TEMP_{MAX.SD}$ is difficult to interpret. That is, the negative correlation of this measure with phonological change cannot be further interpreted regarding language change. The other climate variability measures ($PREC_{SD}$ and $TEMP_{MIN.SD}$), which are more meaningful with regard to Nettle's (1998) theory, correlate according to the expectations. This confirms the appropriateness of the concept behind these measures. A limitation of the climatic factors is that they are likely to change over time (Currie and Mace, 2012). This is partly addressed by using climatic data reflecting the conditions of 6'000 years BP which is more suitable for the time-span that is reflected (10'000 years BP until now). Climatic changes, however, are not incorporated.

The topographical factors generally seem to measure what they are intended to; they show similar correlations within the datasets of phonology and grammar respectively. The correlations among the four topographical measures (ALT_{AV}, ALT_{SD}, TRI_{AV}, TRI_{SD}) suggest a certain stability. Further correlations with temperature measures also point to the appropriateness of the measures: the negative correlation with $TEMP_{MAX.AV}$ is comprehensible due the fact that a rough area has higher elevations leading to cooler temperatures. The positive correlation with temperature variability ($TEMP_{MIN.SD}$) also makes sense because the rougher an area is, the more different elevations there are, which are connected to lower and higher temperatures respectively. The ALT_{AV} measure shows strong correlations, but it is conceptually the most critical measure as it calculates only the average altitude for a region. It is thus not informative for language family areas comprising regions with different topographical characteristics. The ALT_{SD} measure accounts for that, but it does not incorporate neighbourhood in its calculation either, while both TRI measures do that (see section 6.3.2.1). Hence, especially the ALT_{AV} measure has to be interpreted with caution.

8.2.2.3 Neighbourhood Measures

All adjacency measures show the expected results for the family and for the language level and they only show correlations within the neighbourhood measures (see table 6.6 and A.2). For the language level (ADJ_{LANG} measures), this indicates that the distinction between neighbours in total and genetically related and unrelated languages makes sense. This suggests that adjacency is an appropriate measure to assess neighbourhood and thus to quantify contact potential.

The different PD measures show controversial results. Except for a positive tendency of PD_{DF} for grammar on the feature level (very uncertain), all patterns suggest a negative relationship between PD measures and language change. This was expected for the PD_{SF} measures quantifying the number of unrelated language neighbours, but not for the PD_{TOT} and PD_{DF} measures. A possible explanation is that high values of these measures rather show the result of no language contact than contact potential. This is associated with high language diversity (Nettle, 1998), which is indicated by the correlation of the PD measures with $PREC_{AV}$ as well. This suggests that the measures are not suitable for quantifying contact potential.

8.2.2.4 Geometric Properties

The areal size and the perimeter result only in insecure correlations. These geometric measures correlate with several environmental measures as described above. These dependencies suggest that these absolute geometric measures may not be suitable for measuring the probability of effectiveness of language contact. Furthermore, it is unexpected that the areal size and the cardinal size do not show a correlation. This means that big families with regard to cardinal size do not necessarily have a large area. This may be explained by language and language family diversity which is increasing towards the equator.

Based on the correlations with other geographical factors (see table 6.7), the horizontality measures seem to measure what they are intended to: they only correlate negatively with REOCK. This indicates that the more horizontal an area is, the less compact it is, which is not a surprising relationship. The horizontality measures, however, tend to be distorted by outlier languages, which may affect the longitudinal or latitudinal spread heavily. A possible explanation for the weakness of the correlation is that on a smaller scale, climatic conditions do not change considerably. As most of the families do not spread across a large area, horizontality measures are not suitable on the level of language families.

The correlations of compactness measures with other geometric factors suggest that the factors are suitable for quantifying shape compactness of a language family area: the correlation of IPQ with AREA and PERI can be explained by the way the IPQ is computed (see section 6.3.4.4) and REOCK correlates with horizontality, as mentioned above. Based on the resulting correlations with phonological change on the aggregated level, $REOCK_{CORR}$, which accounts for the distinction of landmass and ocean, seems to be more suitable than

REOCK. The simple area-perimeter measure IPQ, however, shows slightly higher correlations than $REOCK_{CORR}$, which indicates that it is suitable as well. This may be due to the fact that both Reock measures are strongly influenced by outlier languages. Moreover, it was expected that the compactness measures would show strong correlations with ADJ_{FAM}. The strongest relationship is observable between the IPQ and ADJ_{FAM} (*rho* = -0.46), which indicates a tendency of compact family areas to have only a few neighbours.

In this chapter the results of the correlation analyses were interpreted with regard to their linguistic meaning and the appropriateness of the geographical factors was reflected on. The next chapter answers the research questions posed in the beginning (section 1.1) and the limitations of the research approach applied in this thesis are elaborated on.

Chapter 9

Discussion

This thesis investigated the influence of geographical factors on contact-induced language change based on the contact hypothesis, which associates fast language change with intense language contact. The contribution of the performed analysis is twofold: on the one hand, new insights are gained into the influence of geography on contact-induced language change. On the other hand, a toolbox providing spherical computation methods for the modelling of geographical characteristics of language family areas has been created. These contributions were guided by two principle research questions outlined in the beginning of this thesis (see section 1.1). In the previous chapter, these questions have already been answered indirectly. In the following, the questions are answered explicitly and methodological and conceptual limitations of the applied research approach are discussed.

9.1 RQ 1: Do geographical factors influence language change?

When answering this and the following questions, it is important to keep in mind that apart from geography, numerous political, socioeconomic and sociolinguistic factors play an important role in contact situations as well. For instance, the intensity of language contact and the social identification with a foreign language are crucial (Thomason, 2001). This principle research question was concretised using two sub-questions, which are answered in the following.

9.1.1 RQ 1.1: Does geography influence grammatical and phonological language change in a similar way?

Different geographical factors, which promote or restrict the emergence of language contact, influence language change. They influence both phonology and grammar. Phonological change, however, shows more correlations with geographical factors suggesting that language contact influences phonology to a higher degree than grammar. The results show that grammatical and phonological change correlate with different geographical factors. Based on the interpretation (chapter 8), these differences are discussed in the following.

Both phonological and grammatical change correlate with climatic conditions, but with different factors. Climate variability correlates only with grammatical change, whereby the correlation is positive. This is in accordance with the hypothesis that the altered climatic conditions a group encounters during migration result in a restructuring process of the social system of a society, which makes contact with other groups necessary (Güldemann, 2010; Diamond, 1997). The positive correlation between phonological change and the average climatic values is strong, suggesting that the warmer the temperature is and the more precipitation occurs in a language family area, the more change in phonology is induced. Regarding language change, these findings do not support Nettle's (1998) ecological risk theory. This theory states that a warm and humid environment promotes self-supply suggesting that a group does not depend on establishing networks to other groups, which in turn results in less language contact between groups.

Topographical complexity correlates with both phonology and grammar, but the orientation of the correlation differs. The hypothesis that topographical complexity favours isolation, which in turn leads to slow language change (Stepp, Castaneda, and Cervone, 2005), is supported for change in phonology but not for change in grammar. As described in section 8.2.2.2, the correlation of topographical complexity and grammatical change is less secure than the correlation of topographical complexity and phonological change. On the one hand, this negative correlation may be ascribed to the underlying data. On the other hand, grammatical change only correlates with more critical geographical factors (ALT_{AV}, ALT_{SD}), which may lead to an overestimation of the influence of topographical complexity on grammar.

Both phonological and grammatical change show dependencies on contact potential, i.e. on the number of adjacent neighbours. The results suggest that it is important for phonological features whether neighbouring languages are related or unrelated. For grammatical change it is only important whether there are a lot of neighbours or not. Both correlations support the hypotheses stating that having a lot of neighbours leads to more change in language and that contact with unrelated neighbours induces more language change than contact with related languages. It is interesting, however, that the dependencies of phonological and grammatical features on contact potential differ. A potential explanation is that grammar may contain less phylogenetic signals than phonology, i.e. phonology of languages within a language family may develop more similarly than their grammar. Hence, although lexical borrowing may happen, it does not leave signals because sound systems of related languages tend to be similar. This is an assumption made by Prof. Dr. Balthasar Bickel during a conversation about this result. To date, however, this has not been tested.

Environmental factors operationalising the probability of contact, and neighbourhood quantifying contact potential, correlate with both phonological and grammatical change. The geometric properties that quantify the probability of contact effectiveness, however, correlate only with phonological change. The hypothesis that stable social structures (indicated by shape compactness) result in less fragmentation and diversification (Diamond,

1997; Trudgill, 2010), which in turn favours resistance against language change in a contact situation, is supported.

Thus, several geographical factors lead to a higher probability of language contact and hence support language change in phonology and grammar. The probability that language contact is effective, however, correlates only with phonological change. The results support the hypothesis that contact, the probability and effectiveness of which is quantified by geographical factors, influences phonological change to a greater degree than grammatical change. This may be attributed to lexical borrowing, which is the most frequent effect of contact and the main carrier of sound changes (Tadmor, 2009; Sankoff, 2002). The borrowing of lexical items often leads to subsequent adjustments in the sound system of a language, which may be applied to the native lexicon as well (Sankoff, 2002).

9.1.2 RQ 1.2: The change of which linguistic features can be explained by geography?

As described in chapter 8, research question 1.2 cannot be answered. In personal conversations with linguists it was discussed that because of too much noise, no signal is observable. This may be due to uncertainties stemming from the phylogenetic trees used for the estimation of the transition rates. In these methods, for each language family the likeliest tree is selected, however, it is not known how accurately these trees depict the evolutionary histories of language families, which are not known in their entirety (Nichols and Warnow, 2008). Additionally, the estimation of transition rates introduces further uncertainties as it is based on these trees. Moreover, only binary data, i.e. features with only two states, and a minimum of ten languages were used for the estimation (see section 5.1). The aggregation of the transition rates to an average value, however, seems to smooth the noise on the feature level, which leads to clearer results indicating the trends described in the previous section.

Answering these questions contributed to gaining a better understanding of the influence of geography on contact-induced language change. The second research question addresses the methodology applied in this thesis.

9.2 RQ 2: How can geographical factors be determined for the investigation of language change?

This thesis has shown that the methodological approach regarding the geographical factors is appropriate. The computed geographical factors are suitable operationalisations of the conceptual factors that are important for contact-induced language change: environmental characteristics (operationalizing the probability of contact), neighbourhood measures (operationalizing contact potential, which is connected to the concept of language probability) and geometric properties (operationalizing the probability of the contact being effective).

An evaluation of the computed geographical factors is not possible, however, based on section 8.2, they can be validated to a certain extent. This is done in the following.

Regarding the climatic measures the factors based on the maximum mean temperature dataset are conceptually critical because they do not influence the productivity of the environment directly, which is crucial for the environmental risk theory by Nettle (1998). For this, the minimum mean temperature is more suitable. The remaining climatic factors seem to be suitable for this scale of analysis. However, the causal relationship between the minimum mean temperature and latitude has to be viewed critically, although, in this thesis, the results are not considerably influenced by it. A limitation of the climatic factors is that climatic and thus also ecological conditions are likely to change over time. Moreover, these factors may influence forager and pastoralist populations more than societies living off agriculture (Currie and Mace, 2012). Including this information into a further analysis may thus be advisable.

The computed topographical measures seem to be suitable. Nevertheless, the simple average height of a region (ALT_{AV}) is conceptually critical, especially for language family areas containing different kinds of topographical elements, which is likely for language families with large areal sizes. This is a conceptual problem stemming from the inequality of language families in terms of cardinal size, areal size, speaker population, etc. This problem is elaborated on in the following limitations section.

Contact potential is best quantified by a definition of neighbourhood based on adjacency. The results derived from these measures are reasonable and the adjacency measures do not show correlations with environmental or geometric factors. The number of neighbours within a certain distance of a language, however, reflects the result of having or not having contact and is thus not a suitable measure.

To measure the probability of the effectiveness of language contact, compactness delivers the most accurate results. Areal size and perimeter were expected to reflect that intense language contact has occurred in case of high values (Currie and Mace, 2009). However, they have shown to be dependent on language and language family density, which is increasing towards the equator. This suggests that these measures are not suitable for measuring the stability of social systems of societies. Compactness, however, is not dependent on the size of an area (i.e. a large area can be as compact as a small area). Thus, the compactness of a language family is not directly influenced by latitude. Horizontality has shown not to be suitable on the granularity of language family areas. A probable explanation is that language family areas are generally too small to reflect migration along the latitudinal axes, contrarily to as was suggested by Diamond (1997). This indicates that horizontality is only suitable for larger scale areas, such as large-scale linguistic areas, which are horizontally aligned if they have large areal sizes, as found by Hammarström and Güldemann (2014).

The approach applied in this thesis shows some limitations, which are addressed in the following section.

9.3 Limitations

Besides some minor limitations such as the drawbacks of the Voronoi method for modelling language family areas (see section 6.2), there are three major limitations of the approach applied: first, the classification and unevenness of language families. Second, uncertainties of the phylogenetic methods applied and third, the concept of representing dynamic processes using static footprints.

The classification of languages into language families entails two limitations: on the one hand, the number of genetic classes is disputed in linguistics and the results of this study would probably turn out differently if another classification were used. On the other hand, the language families are very unequal in cardinal size, areal size, speaker population, etc. This is also reflected in the results of the correlation analyses. For instance, the exclusion of the four biggest families leads generally to stronger correlations on the aggregated level and to more correlations on the feature level. This indicates that the distinction of language families of different cardinal sizes may be useful to yield more meaningful results. Moreover, from a geographical perspective, the greatly varying areal sizes of the language families lead to a problem of scale (see e.g. Goodchild (2001) and Fisher, Wood, and Cheng (2004)). A challenge of this thesis was to find geographical factors that are equally informative for different areal sizes, that is, for a variety of scales.

An aforementioned limitation of the approach applied in this thesis is that it is highly dependent on the transition rate data estimated based on phylogenetic trees. These phylogenetic trees cannot be evaluated for the majority of language families, because they are only well-studied for a few genetic groupings (Nichols and Warnow, 2008). Nevertheless, some linguistic theories are reflected in the results. Consequently, it can be assumed that the phylogenetic methods applied are appropriate despite these limitations.

The biggest drawback of the applied approach is that language evolution, which is a dynamic process, is examined on the basis of today's distribution of languages in space. Today's shape of language family areas is a static footprint of the recent stage of the evolutionary history of languages. The process leading to this stage, however, is not reflected in this footprint. For a more appropriate modelling of language divergence and contact, synchronous data would be necessary. Language evolution would need to be examined together with spatial migration patterns of ethnolinguistic groups. The case of Indo-European illustrates this: Bouckaert et al. (2012) model the expansion of the Indo-European languages through time. Their model shows the expansion from Anatolia starting between 8'000 and 9'500 years ago. Until around 3'000 years BP, the migration took place mainly along latitudinal axes. Only more recent migratory patterns show an expansion into more northern and southern regions of Europe and Asia. This is in line with Diamond's (1997) theory. However, such processes are not reflected in the approach used in this thesis. By using today's distribution of Indo-European languages, this latitudinal direction of the migration is not recorded because of the missing temporal dimension.

Chapter 10

Conclusion

Based on theories originating from research on language diversity, a global analysis of contact-induced language change from a geographical perspective was carried out. The analysis was performed on the scale of language families, i.e. language change was assessed for language families and geographical factors were computed for their respective areas. Appropriate geographical factors were defined to measure the influence of geography on contact-induced language change. Based on these factors, linguistic theories on linguistic diversity were tested and several hypotheses turned out to be supported by the findings.

Contact potential, which is quantified as the number of adjacent neighbours, has a strong influence on change in phonology and grammar. In general, grammatical and phonological change are influenced by different geographical factors. While phonological change is hampered by isolation caused by topographical complexity, grammatical change is favoured in areas with varying climatic conditions. Further, shape compactness of a language family area, which reflects the stability of the social system of societies, correlates negatively with phonological change. This indicates that when language contact occurs, it is less effective if a language family area shows a high compactness and vice versa. The results moreover suggest that phonological change is influenced by geographical factors to a higher degree than grammatical change. This may be attributed to lexical borrowing, which is the most frequent effect of contact and the main carrier of sound change (Tadmor, 2009; Sankoff, 2002).

Chapter 11

Outlook

In the future, the conceptual problem of representing dynamic processes by static footprints may be circumvented by using data from studies based on genomic data. In recent years, the collection of high-resolution human genomic data of different populations worldwide has increased and new methodologies have been developed for inferring population history (see e.g. Li and Durbin (2011)). Such genomic analyses make a contribution to research on migration events (see e.g. Pagani et al. (2016)). The spatiotemporal data resulting from such studies, which show how humans spread around the world, could be linked to linguistic data instead of using the snapshot of today's distribution of languages. This would allow to relate the dynamic system of languages to spatiotemporal migration data.

This thesis has shown which geographical factors are suitable for measuring the influence of geography on contact-induced language change. In a further step, the used geographical factors could be combined with each other in order to statistically model language contact on the basis of geography. The results of this thesis thus serve as a preliminary analysis. Moreover, some of the obtained results cannot be explained by literature, as for example the different correlations of phonological and grammatical change with neighbourhood measures. Such findings could be further investigated, for example on a different (zoomed-in) scale. To be able to draw final conclusions from such findings, further research is required.

Appendix A

Appendix

Table A.1: Abbreviation legend for linguistic features.

Abbreviation	Linguistic feature
affrct..	affricates.p.cat
asprt...	aspirated.fricatives.p.cat
att.flex	autotyp.morphology.per.language$any.flexivity
att.poly	autotyp.morphology.per.language$any.polyexponence
a.$BROHA	autotyp.wals$BROHAN
a.$DRYSO3	autotyp.wals$DRYSOV3
at.$NADJ	autotyp.wals$NADJ
atty.$VP	autotyp.wals$VP
cnstrG..	constrictedGlottis.p.cat
crky.b..	creaky.breathy.p.cat
dstrbt..	distributed.p.cat
lbdntl..	labiodental.p.cat
ltrl.p.c	lateral.p.cat
lqds.p.c	liquids.p.cat
lng.p.ct	long.p.cat
lng.vw..	long.vowels.p.cat
lwrdLI..	loweredLarynxImplosive.p.cat
sprdGl..	spreadGlottis.p.cat
tap.p.ct	tap.p.cat
ton.p.ct	tone.p.cat
trll.p.c	trill.p.cat
vlr.ns..	velar.nasals.p.cat
advnTR..	advancedTongueRoot.p.cat
a.$BROFI	autotyp.wals$BROFIN
a.$DRYNP	autotyp.wals$DRYNPL2
a.$DRYPO	autotyp.wals$DRYPOS2
a.$DRYSO	autotyp.wals$DRYSOV4
atty.$PP	autotyp.wals$PP
bck.p.ct	back.p.cat
clck.p.c	click.p.cat
dlydRl..	delayedRelease.p.cat
frts.p.c	fortis.p.cat
frctvs..	fricatives.p.cat
frnt.p.c	front.p.cat
glds.p.c	glides.p.cat
lqds....	liquids.glides.nasals.p.cat
lqds.g..	liquids.glides.p.cat
low.p.ct	low.p.cat
nsl.p.ct	nasal.p.cat
nsls.p.c	nasals.p.cat
rsdLrE..	raisedLarynxEjective.p.cat
rtrcTR..	retractedTongueRoot.p.cat
rnd.p.ct	round.p.cat
shrt.p.c	short.p.cat
strdnt..	strident.p.cat
tns.p.ct	tense.p.cat
vcls.stp	voiceless.stops.p.cat
vcls.vow	voiceless.vowels.p.cat

Geographical Factors	SIZE	AREA	PERI	HOR_{MAX}	HOR_{MID}	IPQ	REOCK	$REOCK_{CORR}$	ADJ_{FAM}	$ADJ_{LANG.TOT}$	$ADJ_{LANG.SF}$	$ADJ_{LANG.DF}$	$PD_{100.TOT}$	$PD_{500.SF}$	$PD_{100.DF}$	$PD_{500.TOT}$	$PD_{500.SF}$	$PD_{500.DF}$	$PREC_{AV}$	$PREC_{SD}$	$TEMP_{MIN.AV}$	$TEMP_{MIN.SD}$	$TEMP_{MAX.AV}$	$TEMP_{MAX.SD}$	ALT_{SD}	ALT_{AV}	TRI_{AV}	TRI_{SD}
SIZE		0.33	0.42	0.12	0.13	-0.48	-0.09	-0.05	0.41	0.24	0.58	-0.39	0.51	0.53	0.44	0.51	0.73	0.36	0.15	0.47	0.11	0.25	0.10	0.13	0.12	0.06	0.00	0.04
AREA	0.33		0.95	0.15	-0.05	-0.52	-0.06	-0.15	0.42	-0.41	-0.28	-0.15	-0.49	-0.47	-0.41	-0.48	-0.26	-0.57	-0.57	-0.08	-0.54	0.70	0.02	0.64	0.22	0.21	-0.13	-0.12
PERI	0.42	0.95		0.21	0.03	-0.74	-0.21	-0.28	0.49	-0.42	-0.32	-0.13	-0.40	-0.41	-0.28	-0.41	-0.20	-0.47	-0.42	0.03	-0.44	0.69	0.00	0.60	0.18	0.21	-0.10	-0.07
HOR_{MAX}	0.12	0.15	0.21		0.94	-0.24	-0.50	-0.42	-0.15	-0.38	-0.09	-0.27	0.02	0.01	0.07	-0.06	0.03	-0.06	-0.22	-0.16	-0.26	0.09	-0.03	0.18	-0.14	-0.16	-0.23	-0.22
HOR_{MID}	0.13	-0.05	0.03	0.94		-0.19	-0.50	-0.41	-0.16	-0.25	0.02	-0.22	0.18	0.17	0.19	0.10	0.17	0.11	-0.05	-0.04	-0.05	-0.10	0.10	0.01	-0.21	-0.24	-0.26	-0.23
IPQ	-0.48	-0.52	-0.74	-0.24	-0.19		0.51	0.51	-0.46	0.23	0.25	-0.01	0.08	0.12	-0.10	0.10	-0.04	0.08	-0.01	-0.34	0.05	-0.41	-0.08	-0.37	-0.09	-0.12	0.05	-0.01
REOCK	-0.09	-0.06	-0.21	-0.50	-0.50	0.51		0.82	-0.12	0.25	0.34	-0.14	0.01	0.07	-0.19	0.05	0.06	-0.04	-0.07	-0.23	-0.08	-0.15	-0.20	-0.08	0.01	-0.02	0.04	0.01
$REOCK_{CORR}$	-0.05	-0.15	-0.28	-0.42	-0.41	0.51	0.82		-0.19	0.03	0.44	-0.43	0.12	0.22	-0.20	0.06	0.14	-0.06	0.08	0.04	-0.05	-0.02	-0.41	0.04	0.12	0.21	0.34	0.29
ADJ_{FAM}	0.41	0.42	0.49	-0.15	-0.16	-0.46	-0.12	-0.19		0.12	-0.05	0.13	0.02	-0.03	0.17	0.08	0.08	0.05	0.02	0.24	0.10	0.34	0.06	0.27	0.24	0.32	0.04	0.19
$ADJ_{LANG.TOT}$	0.24	-0.41	-0.42	-0.38	-0.25	0.23	0.25	0.03	0.12		0.50	0.42	0.55	0.48	0.62	0.72	0.49	0.73	0.36	0.08	0.42	-0.46	0.29	-0.44	0.01	-0.21	-0.17	-0.14
$ADJ_{LANG.SF}$	0.58	-0.28	-0.32	-0.09	0.02	0.25	0.34	0.44	-0.05	0.50		-0.52	0.80	0.88	0.48	0.76	0.89	0.58	0.38	0.40	0.24	-0.15	-0.10	-0.17	0.13	0.02	0.20	0.19
$ADJ_{LANG.DF}$	-0.39	-0.15	-0.13	-0.27	-0.22	-0.01	-0.14	-0.43	0.13	0.42	-0.52		-0.29	-0.43	0.10	-0.09	-0.44	0.11	-0.03	-0.28	0.18	-0.32	0.45	-0.25	-0.13	-0.25	-0.40	-0.37
$PD_{100.TOT}$	0.51	-0.49	-0.40	0.02	0.18	0.08	0.01	0.12	0.02	0.55	0.80	-0.29		0.97	0.84	0.94	0.88	0.86	0.55	0.47	0.44	-0.30	-0.01	-0.34	0.13	-0.01	0.26	0.26
$PD_{500.SF}$	0.53	-0.47	-0.41	0.01	0.17	0.12	0.07	0.22	-0.03	0.48	0.88	-0.43	0.97		0.71	0.88	0.91	0.76	0.56	0.50	0.43	-0.24	-0.07	-0.28	0.11	0.03	0.30	0.32
$PD_{100.DF}$	0.44	-0.41	-0.28	0.07	0.19	-0.10	-0.19	-0.20	0.17	0.62	0.48	0.10	0.84	0.71		0.89	0.63	0.93	0.47	0.34	0.44	-0.36	0.17	-0.38	0.08	-0.18	0.06	0.02
$PD_{500.TOT}$	0.51	-0.48	-0.41	-0.06	0.10	0.10	0.05	0.06	0.08	0.72	0.76	-0.09	0.94	0.88	0.89		0.85	0.95	0.53	0.39	0.47	-0.39	0.12	-0.43	0.07	-0.14	0.09	0.10
$PD_{500.SF}$	0.73	-0.26	-0.20	0.03	0.17	-0.04	0.06	0.14	0.08	0.49	0.89	-0.44	0.88	0.91	0.63	0.85		0.70	0.45	0.54	0.36	-0.15	0.04	-0.19	0.07	-0.01	0.14	0.18
$PD_{500.DF}$	0.36	-0.57	-0.47	-0.06	0.11	0.08	-0.04	-0.06	0.05	0.73	0.58	0.11	0.86	0.76	0.93	0.95	0.70		0.53	0.31	0.52	-0.50	0.22	-0.49	0.01	-0.21	0.04	0.03
$PREC_{AV}$	0.15	-0.57	-0.42	-0.22	-0.05	-0.01	-0.07	0.08	0.02	0.36	0.38	-0.03	0.55	0.56	0.47	0.53	0.45	0.53		0.56	0.84	-0.42	-0.12	-0.59	-0.31	-0.13	0.16	0.22
$PREC_{SD}$	0.47	-0.08	0.03	-0.16	-0.04	-0.34	-0.23	0.04	0.24	0.08	0.40	-0.28	0.47	0.50	0.34	0.39	0.54	0.31	0.56		0.39	0.12	-0.02	0.06	0.11	0.30	0.36	0.39
$TEMP_{MIN.AV}$	0.11	-0.54	-0.44	-0.26	-0.05	0.05	-0.08	-0.05	0.10	0.42	0.24	0.18	0.44	0.43	0.44	0.47	0.36	0.52	0.84	0.39		-0.57	0.26	-0.66	-0.48	-0.28	-0.09	0.04
$TEMP_{MIN.SD}$	0.25	0.70	0.69	0.09	-0.10	-0.41	-0.15	-0.02	0.34	-0.46	-0.15	-0.32	-0.30	-0.24	-0.36	-0.39	-0.15	-0.50	-0.42	0.12	-0.57		-0.41	0.83	0.56	0.71	0.43	0.42
$TEMP_{MAX.AV}$	0.10	0.02	0.00	-0.03	0.10	-0.08	-0.20	-0.41	0.06	0.29	-0.10	0.45	-0.01	-0.07	0.17	0.12	0.04	0.22	-0.12	-0.02	0.26	-0.41		-0.21	-0.34	-0.51	-0.63	-0.55
$TEMP_{MAX.SD}$	0.13	0.64	0.60	0.18	0.01	-0.37	-0.08	0.04	0.27	-0.44	-0.17	-0.25	-0.34	-0.28	-0.38	-0.43	-0.19	-0.49	-0.59	0.06	-0.66	0.83	-0.21		0.59	0.65	0.35	0.36
ALT_{AV}	0.12	0.22	0.18	-0.14	-0.21	-0.09	0.01	0.12	0.24	0.01	0.13	-0.13	0.13	0.11	0.08	0.07	0.07	0.01	-0.31	0.11	-0.48	0.56	-0.34	0.59		0.71	0.64	0.53
ALT_{SD}	0.06	0.21	0.21	-0.16	-0.24	-0.12	-0.02	0.21	0.32	-0.21	0.02	-0.25	-0.01	0.03	-0.18	-0.14	-0.01	-0.21	-0.13	0.30	-0.28	0.71	-0.51	0.65	0.71		0.77	0.84
TRI_{AV}	0.00	-0.13	-0.10	-0.23	-0.26	0.05	0.04	0.34	0.04	-0.17	0.20	-0.40	0.26	0.30	0.06	0.09	0.14	0.04	0.16	0.36	-0.09	0.43	-0.63	0.35	0.64	0.77		0.92
TRI_{SD}	0.04	-0.12	-0.07	-0.22	-0.23	-0.01	0.01	0.29	0.19	-0.14	0.19	-0.37	0.26	0.32	0.02	0.10	0.18	0.03	0.22	0.39	0.04	0.42	-0.55	0.36	0.53	0.84	0.92	

Table A.2: Correlation matrix of geographical factors.

Bibliography

Amante, C. and B.W. Eakins (2009). *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. URL: https://www.ngdc.noaa.gov/mgg/global/global.html.

Ascione, A. et al. (2008). "The Plio-Quaternary Uplift of the Apennine Chain: New Data from the Analysis of Topography and River Valleys in Central Italy". In: *Geomorphology* 102.1, pp. 105–118.

Atkinson, Quentin D. (2011). "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa". In: *Science* 332, pp. 346–349.

Atkinson, Quentin D and Russell D Gray (2005). "Curious Parallels and Curious Connections: Phylogenetic Thinking in Biology and Historical Linguistics". In: *Systematic Biology* 54.4, pp. 513–526.

Atkinson, Quentin D. et al. (2008). "Languages Evolve in Punctuational Bursts". In: *Science* 319.5863, p. 588.

Aurenhammer, Franz (1991). "Voronoi Diagrams - A Survey of a Fundamental Data Structure". In: *ACM Computing Surveys* 23.3, pp. 345–405.

Bickel, Balthasar (2015). "Distributional Typology: Statistical Inquiries into the Dynamics of Linguistic Diversity". In: *The Oxford Handbook of Linguistic Analysis*. Ed. by Bernd Heine and Heiko Narrog. 2nd ed. Oxford: Oxford University Press. Chap. 37, pp. 901–923.

— (2016). *Transition Rate Estimates Across Families and Variables. R Script.*

Bickel, Balthasar and Johanna Nichols (in prep.). *The AUTOTYP database. Electronic database.*

Bouckaert, Remco et al. (2012). "Mapping the Origins and Expansion of the Indo-European Language Family". In: *Science* 337.August, pp. 957–960.

Bowern, Claire (2013). "Relatedness as a Factor in Language Contact". In: *Journal of Language Contact* 6, pp. 411–432.

Bowern, Claire and Bethwyn Evans, eds. (2015). *The Routledge Handbook of Historical Linguistics*. London: Routledge, p. 757.

Bromham, Lindell et al. (2015). "Rate of Language Evolution is Affected by Population Size". In: *Proceedings of the National Academy of Sciences* 112.7, pp. 2097–2102.

Collard, Ian F and Robert A Foley (2002). "Latitudinal Patterns and Environmental Determinants of Recent Human Cultural Diversity: Do Humans Follow Biogeographical Rules ?" In: *Evolutionary Ecology Research* 4.3, pp. 371–383.

Cormen, Thomas H. et al. (2009). *Introduction to Algorithms (3rd ed.), Cambridge, MA: MIT Press: 1029-1034*. Cambridge.

Cox, E.P. (1927). "A Method of Assigning Numerical and Percentage Values to the Degree of Roundness of Sand Grains". In: *Journal of Paleontology* 1.3, pp. 179–183.

Currie, Thomas E and Ruth Mace (2009). "Political Complexity Predicts the Spread of Ethnolinguistic Groups". In: *Proceedings of the National Academy of Sciences* 106.18, pp. 7339–7344.

Currie, Thomas E. and Ruth Mace (2012). "The Evolution of Ethnolinguistic Diversity". In: *Advances in Complex Systems* 15.1 & 2.

Dediu, Dan (2015). *lgfam-Newick: Language Family Classifications as Newick Trees*. URL: https:{\%}0A//github.com/ddediu/lgfam-newick.

Diamond, J. (1997). *Guns, Germs and Steel: The Fates of Human Societies*. London: Cape, p. 425.

Dixon, Robert M. W. (1997). *The Rise and Fall of Languages*. Cambridge: Cambridge University Press, p. 175.

Donohue, Mark et al. (2013). *World Phonotactics Database*. Canberra: Department of Linguistics, The Australian National University. URL: http://phonotactics.anu.edu.au.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. URL: http://wals.info/.

Duckham, Matt et al. (2008). "Efficient Generation of Simple Polygons for Characterizing the Shape of a Set of Points in the Plane". In: *Pattern Recognition* 41.10, pp. 3224–3236.

Dunn, Michael (2015). "Language Phylogenies". In: *The Routledge Handbook of Historical Linguistics*. Ed. by Claire Bowern and Bethwyn Evans. London: Routledge. Chap. 7, pp. 190–211.

Enfield, N.J. (2005). "Areal Linguistics and Mainland Southeast Asia". In: *Annual Review of Anthropology* 34.1, pp. 181–206.

ESRI (2014). *Natural Earth Dataset 10m*. URL: http://www.naturalearthdata.com/.

Fisher, Peter, Jo Wood, and Tao Cheng (2004). "Where is Helvellyn? Fuzziness of Multi-Scale Landscape Morphometry". In: *Transactions of the Institute of British Geographers* 29.1, pp. 106–128.

Gavin, Michael C. et al. (2013). "Toward a Mechanistic Understanding of Linguistic Diversity". In: *BioScience* 63.7, pp. 524–535.

Goodchild, M.F. (2001). "Models of Scale and Scales of Modelling". In: *Modelling Scale in Geographical Information Science*. Ed. by N. J. Tate and P. M. Atkinson. New York: John Wiley & Sons, pp. 3–10.

Greenhill, S. J. et al. (2010). "The Shape and Tempo of Language Evolution". In: *Proceedings of the Royal Society of London B: Biological Sciences* 277.1693, pp. 2443–2450.

Güldemann, Tom (2010). ""Sprachraum" and Geography : Linguistic Macro-Areas in Africa". In: *Language and Space: An International Handbook of Linguistic Variation: Language Mapping*. Ed. by Alfred Lameli, Roland Kehrein, and Stefan Rabanus. Berlin: Mouton de Gruyter. Chap. 29, pp. 561–585.

Hammarström, Harald (2010). "A Full-Scale Test of the Language Farming Dispersal Hypothesis". In: *Diachronica* 27.2, pp. 197–213.

Hammarström, Harald and Tom Güldemann (2014). "Quantifying Geographical Determinants of Large-Scale Distributions of Linguistic Features". In: *Language Dynamics and Change* 4.1, pp. 87–115.

Hammarström, Harald et al., eds. (2016). *Glottolog 2.7*. Jena: Max Planck Institute for the Science of Human History. URL: http://glottolog.org/.

Heggarty, Paul (2006). "Interdisciplinary Indiscipline ? Can Phylogenetic Methods Meaningfully Be Applied to Language Data — and to Dating Language ?" In: *Phylogenetic Methods and the Prehistory of Languages*. Ed. by Peter Forster and Colin Renfrew. Cambridge: McDonald Institute for Archaeological Research. Chap. 16, pp. 183–194.

Isaaks, Edward H. and R. Mohan Srivastava (1989). *Introduction to Applied Geostatistics*. New York: Oxford University Press, p. 561.

Kaye, Alan S. and Mauro Tosco (2003). *Pidgin and Creole Languages: A Basic Introduction*. Munich: Lincom Europa, p. 113.

Kerr, Jeremy T. and Laurence Packer (1997). "Habitat Heterogeneity as a Determinant of Mammal Species Richness in High-Enery Regions". In: *Nature* 385.6613, pp. 252–254.

Köhli, Martina (2013). "Quantitative Analyse des Zusammenhangs zwischen der globalen Sprachendiversität und geographischen Faktoren". PhD thesis. University of Zurich, p. 129.

Li, Heng and Richard Durbin (2011). "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357, pp. 493–496.

Li, Wenwen et al. (2014). "NMMI : A Mass Compactness Measure for Spatial Pattern Analysis of Areal Features". In: *Annals of the Association of American Geographers ISSN:* 104.6, pp. 1116–1133.

Lucas, Christopher (2014). "Contact-Induced Language Change". In: *The Routledge Handbook of Historical Linguistics*. Ed. by Claire Bowern and Bethwyn Evans. London: Routledge. Chap. 24, pp. 519–536.

McGregor, William B. (2015). *Linguistics: An Introduction*. 2nd. London: Bloomsbury Academic, p. 496.

Meade, Andrew and Mark Pagel (2014). *BayesTraits V2. Manual*. URL: http://www.evolution.rdg.ac.uk/BayesTraitsV2.0Files/TraitsV2Manual.pdf.

Meteorological Research Institute (MRI) (2016). *MRI-CGCM3. Paleoclimate Data*. URL: http://www.worldclim.org/paleo-climate1.

Montero, Raul S and Ernesto Bribiesca (2009). "State of the Art of Compactness and Circularity Measures". In: *International Mathematical Forum* 4.27, pp. 1305–1335.

Moore, Joslin L et al. (2002). "The Distribution of Cultural and Biological Diversity in Africa". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269.1501, pp. 1645–1653.

Moran, Steven, Daniel McCloy, and Richard Wright, eds. (2014). *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: http://phoible.org/.

Nettle, Daniel (1996). "Language Diversity in West Africa : An Ecological Approach". In: *Journal of Anthropological Archaeology* 15, pp. 403–438.

— (1998). "Explaining Global Patterns of Language Diversity". In: *Journal of Anthropological Archaeology* 17.4, pp. 354–374.

Nichols, Johanna (1990). "Linguistic Diversity and the First Settlement of the New World". In: *Language* 66.3, pp. 475–521.

— (1992). *Linguistic Diversity in Space and Time*. Chicago and London: University of Chicago Press, p. 358.

— (2014). "The Vertical Archipelago: Adding the Third Dimension to Linguistic Geography". In: *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*. Ed. by P. Auer et al. Berlin: Walter de Gruyter, pp. 38–60.

Nichols, Johanna and Tandy Warnow (2008). "Tutorial on Computational Linguistic Phylogeny". In: *Language and Linguistics Compass* 2.5, pp. 760–820.

Nichols, Johanna et al. (in prep.). *The AUTOTYP Typological Database, Version 0.1.0*. Zurich: University of Zurich (to be released via GitHub in February 2017).

Pagani, Luca et al. (2016). "Genomic Analyses Inform on Migration Events During the Peopling of Eurasia". In: *Nature* 538.7624, pp. 238–242.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna. URL: https://www.r-project.org/.

Reock, Ernest (1961). "A Note: Measuring Compactness as a Requirement of Legislative Apportionment". In: *Midwest Journal of Political Science* 5.1, pp. 70–74.

Riley, Shawn J, Stephen D DeGloria, and Robert Elliot (1999). "A Terrain Ruggedness Index that Qauntifies Topographic Heterogeneity". In: *Intermountain Journal of Sciences* 5.1-4, pp. 23–27.

Sankoff, Gillian (2002). "Linguistic Outcomes of Language Contact". In: ed. by Peter Trudgill, J. Chambers, and N. Schilling-Estes. London: Blackwell. Chap. 25, pp. 638–668.

Scrucca, Luca (2013). "GA: A Package for Genetic Algorithms in R". In: *Journal of Statistical Software* 53.4, pp. 1–37.

Stepp, John Richard, Hector Castaneda, and Sarah Cervone (2005). "Mountains and Biocultural Diversity". In: *Mountain Research and Development* 25.3, pp. 223–227.

Sujoldžić, Anita and Vesna Muhvić-Dimanovski (2004). "Language Dynamics and Change: Introduction to Linguistic Diversity in Anthropological Perspective". In: *Collegium Antropologicum* 28.1, pp. 1–4.

Tadmor, Uri (2009). "Loanwords in the World's Languages : Findings and Results". In: *Loanwords in the World's Languages. A Comparative Handbook*. Ed. by Martin Haspelmath and Uri Tadmor. Berlin, Boston: De Gruyter Mouton. Chap. 3, pp. 55–75.

Thomason, Sarah (2010). "Contact Explanations in Linguistics". In: *The Handbook of Language Contact*. Ed. by Raymond Hickey. John Wiley & Sons. Chap. 1, pp. 1–12.

Thomason, Sarah G. (2001). *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press, p. 310.

Thomason, Saray G. and Terrence Kaufman (1988). *Language contact, creolization and genetic linguistics*. Berkeley: University Press of California, p. 428.

Trudgill, Peter (2010). "Social Structure and Change". In: *Sociolinguistics Handbook*. Ed. by Ruth Wodak, Barbara Johnstone, and Paul E. Kerswill. London: SAGE. Chap. 17, pp. 236–248.

— (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press, p. 236.

Wichmann, Søren et al. (2015). *The ASJP Database (Version 16)*. URL: `http://asjp.clld.org`.

Widmer, Manuel et al. (2016). "NP Recursion Over Time. Ms. under review". PhD thesis.

Zeige, Lars Erik (2015). "Word forms , Classification, and Tamily Trees of Languages - Why Morphology is Crucial for Linguistics". In: *Zoologischer Anzeiger* 256, pp. 42–53.

Declaration of Authorship

I hereby declare that the material contained in this thesis is my own original work. Any quotation or paraphrase in this thesis from the published or unpublished work of another individual or institution has been duly acknowledged. I have not submitted this thesis, or any part of it, previously to any institution for assessment purposes.

Zurich, January 27th 2017

Fabiola Kälin