University of Zurich[UZH]

Department of Geography

The Impact of Geographical Isolation on Language Particularity

A GIS approach to model Accessibility

GEO 620 Master's Thesis

Author

Hella Mönkeberg

12-721-080

Supervised by

Dr. Curding Derungs

Co-Supervised by

Dr. Rik van Gijn

Faculty representative

Prof. Dr. Robert Weibel

30.06.2018

Department of Geography, University of Zurich

Abstract

Contact-induced language change is a major component of language development. Importantly, the contact of languages is highly dependent on the environment and geography. Thus, the main assumption of this thesis is that languages that are geographically isolated are positively correlating to language particularity. To date, the concept and implementation of geographic isolation are kept very simple. Given the lack of research, this thesis is meant to fill a void and present a model as a realistic approach to defining isolation with regards to reachability in the manner of historical human movement. The basics of the model comprise an accumulative cost surface that allows the inclusion of different information layers. Within this thesis, the model has been fed with data from the elevation model and the historical land surface in the Americas. With the help of this model, the impact of geographical reachability to *language particularity* has been determined. Therefore, quantitative measurements of language particularity due to i. phylogenetic relations and ii. typological properties have been developed. In addition, it will be shown that such indices are highly sensitive to their scaling-level. This insight opens up the analysis on spatial distribution of language traits.

Acknowledgments

An exciting time comes to a close by submitting this master thesis. I would like to express my special thanks to the following people:

- Dr. Curdin Derungs, main supervisor, for your super guidance, communication and professional support,
- Dr. Rik van Gijn, co-supervisor, for introducing me into the field of linguistics and the support,
- Fabian Mönkeberg, for your critical review and all the great encouragement,
- Valeria Pittini, for your professional and prompt proofreading,
- Rafaela Catena, for your proofreading and great assistance,
- Christine Albrecht, for your proofreading and the relaxing times doing sports,
- Elena Votik, for your proofreading and having an open ear,
- Department of Geography for offering the room G10.

Hella Mönkeberg

Table of Contents

List of Figures

List of Tables

1 Introduction

The present linguistic landscape has evolved over approximately two hundred thousand years, stretching out over the globe. Therefore, languages are phenomena with both a spatial and temporal dimension. The field of historical linguistics addresses the history of languages and considers languages as a *window to the past* (Heggarty 2015, Thomason 2001). Historical events, migrations, societies and the environment have formed the languages that are spoken today. As such, historical linguistics is closely connected to the field of archeology. Archeology will be treated broadly as the study of historical human behavior. The insights about language development help to understand historical events. Reciprocally, the more insights are gained through archaeology about historical population and environment, the more conclusions can be drawn in terms of the development of languages. This continuous and important exchange between the disciplines is visualized in Figure 1. Moreover, the visualization shows that geographic information systems (GIS) provide meaningful tools for both fields. While archeology has been using GIS since the 80s, the use of quantitative approaches within the field of linguistics has come up only recently (Heggarty 2015, Howey 2011) – with the exception of the area of dialect studies.

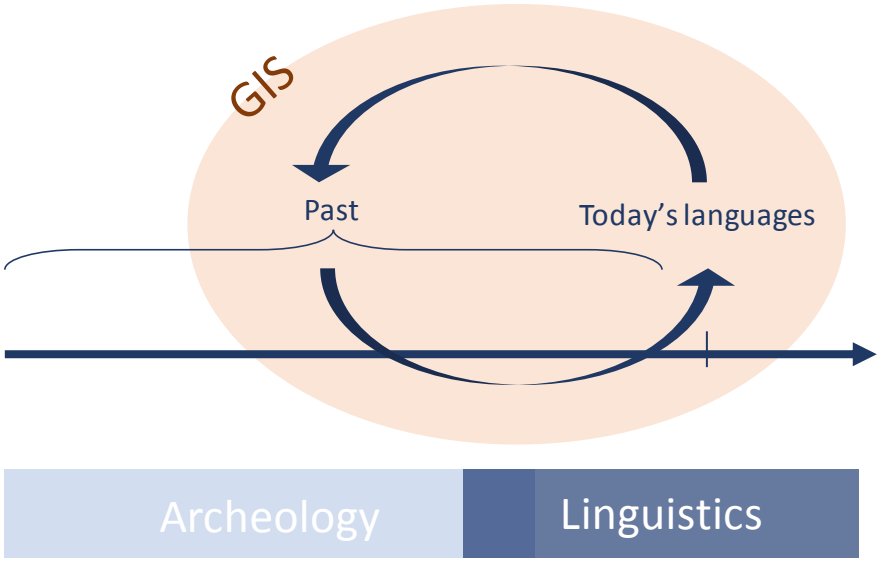


Figure 1: Visualization of the relation between languages - studied in linguistics, the past - studied in Archeology and GIS as a tool in both fields

Language change is a major subject within the field of historical linguistics (Thomason 2008). Languages have always been undergoing continuous change, which has given rise to the awe-inspiring variety we find today (Aitchison 2005). The change of languages can be classified into two

main processes. The first process relates to the contact between speakers of different communities and seems to have the major influence on the development of languages (Thomason 2008). In scientific studies, this phenomenon is called contact-induced language change and can be explained as "speakers of two (or more) languages need to be in the same place for language contact to occur" (Thomason 2001). Linguists do not only acknowledge the existence of contact-induced change, they furthermore agree on the complexity of the process of language transformation. Consequently, it is still not clear how this change can be quantified or how contact-induced language change can be grasped in detail (Thomason 2001).

A generally low degree of contact can be defined as isolation. The process of change in for example geographically isolated languages provides an interesting perspective on the second type of language change known as internally-caused change. In other words, this change relates mainly to imperfect language transmission from parent to child. To give an example, linguists explain the conservative characteristics of Icelandic compared to other Nordic languages as the result of the geographical isolation of the island. The lack of contact of the Icelandic population to other languages has led to a slower rate of change, mainly caused by imperfect transmission (Lucas 2015).

The term *geographical isolation* to explain Icelandic underlines two very important underlying properties of languages. First, languages are a geographical phenomenon over which space and time have crucial influence. Second, the degree of accessibility of a community seems very important for the development of languages. Therefore, accessibility must be understood in terms of human movement. Through human movement, communities and languages have come into contact and have exchanged knowledge and behavior. Historical human movement has mostly been studied within the field of archeology, but spatio-temporal analysis and methods are common tools in geographic information science (GIS). The combination of linguistics, archeology and GIS opens up new opportunities to get insights on the spatial dependency of languages (Heggarty 2015).

Through the cooperation of the three scientific fields, new insights about geographical dependencies of languages have been gained. Some investigations have been able to show that languages that are geographically close to each other are more similar than languages that are geographically distant (Holman et al. 2007). Other studies have been able to show the positive correlation between language complexity and latitude or language diversity and the amount of rainfall or, more generally speaking, the crop growth (Nichols 2013, Nettle 1999). Rogers et al. (1990) proposes explanations of today's languages based on historical environment. These examples show the high impact of geography on language development. However, the metric

distance, the change in altitude or the amount of rainfall are probably not the direct causes for language properties. More likely, they are indices for the degree of isolation, respectively contact to other communities. Today's research shows a clear lack of realistic and suitable models on isolation in terms of human movement and language development.

The examples named above share the use of quantitative analysis. Archaeology and GIS provide different approaches to model the accessibility of a landscape. The most common approach is the least-cost-path analysis (Holmann et al. 2007, Anderson & Gillam 2000, Howey 2007), in which the costs of moving from one point to another within a terrain are computed and the route of the least costs defined as the best route choice. Unfortunately, it is unrealistic to assume any known destination in human movement. Additionally, most models only consider topography as the single factor in human movement. To be able to combine language characteristics and human movement analysis, new methods and approaches must be developed.

Overall, this thesis aims to explain *language particularity* due to geographical accessibility. Language particularity describes the relative uniqueness of a language. To date, the quantitative measurement of language particularity has only rarely been addressed in linguistics. Thus, in a first step, this work determines some appropriate methods to measure language particularity and presents the advantages and disadvantages of different approaches. In geography, the discussion about scaling dependencies is very common. Therefore, within the cooperation of linguistics and geography, it seems obvious to determine the influence of different scaling-levels to the measurement of language particularity. The second part of this thesis focuses on the modeling of geographical accessibility. Thereby, the concept of *isolated-by-distance (IBD)* will be developed so that accessibility can be modeled more realistically in terms of human movement. Based on these two results of language particularity and the degree of reachability, this thesis aims to answer the research question whether the level of accessibility that is shaped by environmental factors can explain language particularity. A more detailed explanation and some further research questions are presented in chapter 4. This analysis will be conducted with regards to the indigenous languages of the Americas.

The next two chapters will give an overview on the state of language development and the geographical dependencies of languages. The lack of research is then presented in chapter 4, which also explains the research questions and hypothesis of this master thesis. Afterwards, the methodologies and approaches to answer these questions are presented. The ensuing chapter summarizes all the important results that are discussed and interpreted in chapter 7.

2 Languages

This master thesis deals with the development and geographical dependencies of languages. Therefore, the main structures and properties of languages are briefly summarized in this section. The first part of *1 Languages* presents an overview of the state of languages and the development of languages. Thereby, the main theories on language change, that are internally-driven and contact-induced language change, will be shortly explained. Moreover, some approaches and methods to measure language particularity will be discussed. To conclude this chapter, the last section provides an outline of the study area of the Americas.

2.1 Overview

Around 7000 languages are spoken around the world. If two people can communicate with each other without learning the language of their conversation partner, they might speak different dialects, but the same language. If they cannot understand each other, one can assume the existence of two different languages. That being said, the differentiation between languages and dialects is gradual and indistinct (Thomason 2001). The distribution of the languages around the world is shaped by human history. Dating back to the first homo sapiens sapiens in Africa and their subsequent expansion around the world, languages have emerged over many thousands of years. Linguists can trace back genealogical relations to about 8,000, but some languages have either split off so long ago, or else have changed so fast, that no genealogical relationship can be established with another language (Nichols 1992). These languages are called *isolate languages*. For other languages, genetic relations could be determined and traced back to a proto-language. If languages descend from a common proto-language, it means that they belong to the same *language family*. Different methods exist to distinguish related languages. Often, this is done by comparing vocabulary. Due to the different approaches, it is not surprising to get different outcomes of family classifications. However, there are some related languages or language families that linguists largely agree on and others that give rise to controversial debates (McGregor 2015).

This thesis is about comparing languages. Along these lines, the following section briefly describes the structure and properties of language. To achieve that, the most important subsystems of languages are explained. If it is not referenced differently, the information is taken from McGregor (2015).

Phonetics and phonology

Phonetics describes the production and perception of language sounds. Different movements of the mouth and tongue produce different airways that can be classified in *phones.* A first major class of phones are *consonants,* which are defined by the place and manner in which air stream is manipulated. For instance, dental consonants are formed by blocking the air stream with the tongue and the upper teeth. A second major class of phones are vowels, which are differentiated according to the location of production within the mouth, such as high, mid and low. Phonology studies the behavior of contrastive sounds and prosody.

Morphology

Morphology deals with the structure of segments within words (Aitchison 2005). A word may have internal complexity, e.g. when a *minimal free form* like 'farm' is combined with an affix such as '-er' in 'farmer'. Morphology describes the principles, constraints, and rules that govern how word-internal segments may be combined.

Lexicon

The lexicon can be visualized as a mental dictionary. It treats the meaning and categorization of all words. The categories are divided into nouns, verbs, adjectives and so on. The lexicon experiences a continuous change as of for example new words or meanings enter a language.

Syntax

Syntax describes the rules to combine words into a sequence or sentence. Sentences are generally structured hierarchically and form words into groups.

Meaning

Meaning consists of semantics and pragmatics. Whereas semantics defines the meaning of the sentence or the word itself, pragmatics is about the inferred information. In other words, pragmatics describes the interpretation and actual understanding of the sentence.

2.1.1 Typology

So far, the main structure of languages has been described. Typology studies the ways in which languages are similar or different with respect to individual properties that relate to language modules described above. These properties (or variables) are usually referred to as *features* in typology. Features allow to compare languages and to define similarities or dissimilarities. The variation in which languages can differ from each other is immense (Cysouw 2011). In general,

languages can differ in each subsystem. For example, they can be distinguished according to the number of vowels, number of phonemes or number of cases. Based on specific kinds of features, certain typological groups of languages can be defined such as the click-languages. However, there is a lot of disagreement about such typological groupings and different literature has offered contradictory findings. Ideally, all languages around the world would be categorized for all existing features (McGregor 2015). Unfortunately, most of the languages are only poorly described. Chapter 2.5 discusses the limitations that stem from this lack of analysis.

2.2 Language Change

Just like everything else, languages change over time (Aitchison 2005). Thereby, the controversial issue is speed and the underlying factors that drive language change. The major agreement is that there are a lot of interdependencies that form a continuous process. Within 400 years, one language can massively change, whereas others alter only little in twice the amount of time (McGregor 2015). Language change can be divided into two main processes, which are internally-driven and contact-induced change, are explained in the following sections.

2.2.1 Internally-driven Language Change

One reason for language change is to be found in the tendency of human beings to streamline and standardize parts of language, especially with regards to the elements that are frequently used. This may lead to a reduction of particular structures or constructions in languages. Another cause of language development is the emergence of new needs and purposes or the necessity to strengthen group identity. Interestingly, it is rare that new words are invented totally independently within a language. This process is known as Coinage (McGregor 2015).

Adjustments of that nature are known as internally-driven language change, that is, in general, a rather slow process and a rarely observed one, too. Nevertheless, the importance does not lie in the rate of change but in the differences in development due to isolation. The next section describes the much faster process of contact-induced language change.

2.2.2 Contact-induced Language Change

Usually, languages are transmitted from parents to children. This is the reason why genetics are often considered within language analysis (Heggarty 2015). However, languages can also be learned through contact with other people and thus other languages. In general, languages that have been in contact become more similar to each other (Aikhenvald 2003). The question of language change

due to contact between communities represents the most important component of historical linguistics (Thomason 2008).

Language contact can be described as the 'use of more than one language in the same place at the same time' (Thomason 2001). All linguistic subsystems can be affected through contact-induced language change. Still, there is no general agreement on this process. While there are studies that claim the existence of non-changeable features, others are concluding that any property of a language can change during such contact (Thomason 2008).

However, the process of adapting and changing is always dependent on the origin language and the preliminary conditions of its community. Circumstances with regards to population size and typological distances of the coinciding languages often influence outcomes of language contact. The degree of language change is highly dependent on the intensity and duration of the contact. The size of the communities and the socioeconomic dominance influence which of the languages is likely to change to a higher degree (Thomason 2001).

There are different types of contacts that result from different types of situations. If a speaker of language A learns language B, some features are likely to be taken over from language A – consciously or unconsciously. This phenomenon is known as substratum interference (Thomason 2008). The other way around, the process that describes when a native speaker uses some expression from a foreign language is called borrowing (Thomason 2008). Borrowing is the most common way to add new words to a language. Specifically, words are defined as loanwords. The number of loanwords for English is assumed to be around 75% of all words (Thomason 2001). In the process of borrowing and after some time, phonology is often adapted to the rules of the receiving language (McGregor 2015). The most extreme result of language contact is language death. This can either be caused by an extreme language shift by the speakers or through the disappearance of a whole community due to a massacre or foreign diseases (Thomason 2001).

In sum, languages change the most and the fastest as a result of contact and, importantly, languages that are in contact are likely to converge (Bowern 2013). Consequently, it can be assumed that the more isolated a language is, the more likely it is to develop in different ways compared to other languages, i.e. the more particular a language becomes overall. Thus, the most particular languages are assumed to be the most isolated. With regards to language development, the degree of contact and also the degree of isolation can be decisive indicators for language particularity. To test this assumption in a quantitative analysis, two measurements are needed. The first one is a quantitative or statistical measurement of language similarity and therefore particularity, which will be discussed

in the next chapter. With respect to the second measurement, there is the need to define the degree of isolation or contact for languages. Chapter 3 presents the conceptual difficulties and Chapter 5.2 highlights the technical challenges of such models.

2.3 Language Particularity

Language particularity is not a commonly used expression in linguistics. In recent studies, similar concepts have been described as similarity and dissimilarity, and their unit of measure are diversity or rarity (Cysouw 2011, Georgi et al. 2010, Nichols 2013). All these concepts describe to which degree languages do or do not differ from each other. To date, linguistic research has mostly focused on the similarity of languages instead of their dissimilarity (Aikhenvald & Dixon 1992). Within this master thesis, the term of *particularity* is mainly used to assess the characteristics of a language in comparison to others. With this in mind, the expressions used in literature can be used in the same manner. One reason to change the expressions is that most of the terms such as similarity or diversity have been used relatively randomly and without determinations or explanations to qualify them further. This master thesis is about_showing the importance of carefully discussing and analyzing this concept. So, to change the expressions also implies a change of the paradigm in which they are used.

In recent years, large language databases have been created that allow quantitative analysis of languages and their properties. The establishment of quantitative analysis offers new possibilities for linguistic research (Cysouw 2013). As further explained in chapter 5, this thesis focuses on methods to compare languages and to find a measurement to define language particularity. Due to the fact that this is a rather new field of research, no consensus about methods and processes has been found so far. However, there are two main approaches to describe language particularity in large datasets, which will be presented in the next sections.

2.3.1 Genealogical particularity

For many linguists, statements about similarities or dissimilarities between languages are based on their family affiliation (Dahl 2008). Many linguists claim that genealogically related languages are more similar to each other than unrelated languages. This leads to the assumption that the smaller the language family, the more particular a language may be. Using this approach, particularity measurements can be computed very easily. As pointed out in section 2.1, the determination of phylogenetic relations is mostly based on vocabulary properties (McGregor 2015).

In spite of what has just been illustrated, there are two insights that cast doubt on this method. The first one is that not all linguists agree on the classification of languages into their families. Especially with regards to the Americas, large discrepancies can be found. While some linguists have only defined three different languages, others have found 57 different families in the Americas (Greenberg 1987, Dahl 2008). In accordance with these disagreements, particularity measurements would show very large discrepancies depending on the input data. The second problem that presents itself when using genealogical particularity is that some scientists do not support the higher degree of similarity of related languages. In fact, they claim that languages of different families are more likely to share more similar typology than related languages (Georgi et al. 2010).

2.3.2 Typological Particularity

The assumption that languages are more similar if they share the same family sounds reasonable, but different studies have shown that languages of different families can be more similar than languages within the same family. This conclusion is based on typological analyses. In other words, it has been shown that typological similarity does not correlate with family membership. In conclusion, typological similarity is also dependent on geographical closeness, and not only on family membership (Georgi et al. 2010).

Most of the typological measurements include different features of all subsystems. This is a very important difference to family membership where mainly vocabulary is compared. With the rise of quantitative datasets, typological comparison and analysis have evolved. The more features are systematically defined for each language, the better languages can be compared to each other. However, as the use of large datasets in linguistics has only started in recent years, the datasets have only sparse information, which is highly problematic. The qualitative field work that goes into gathering all this information is very time consuming. Moreover, scientists also run the risk of subjective perceptions and of categorizations of languages. In sum, the risks and uncertainties are manifold.

This thesis will determine measurements of typological particularity and question the results in terms of usability and improvements. Due to the disagreement on the two methods, both approaches of genealogical as well as typological particularity are used for this analysis.

2.3.3 Relative Measurement

As already pointed out, there is no general agreement on the exact term to define language particularity. Nevertheless, most scientists are using the concept as a natural and clear

measurement. So far, no literature has been found that questions and discusses the meaning of this quantitative language measurement.

The two indices of genealogical and typological particularity are results of comparative statistics. This means that the index is always a measurement in relation to other languages. As a logical consequence, when changing the sample set, the language particularity of each language is likely to change, too. As a result, it is not possible to define an absolute value of rarity, but rather, it is always about a value in relation to a specific sampling.

This has been visualized in Figure 2. The colors represent language characteristics within a specific area. In the figure on the left hand side, the languages within the blue area represent the minority and are thus more particular than the green ones. The image on the right illustrates a higher level of scale as a result of zooming in. Based on the new sample set, the blue languages represent the common and the green languages the particular ones.

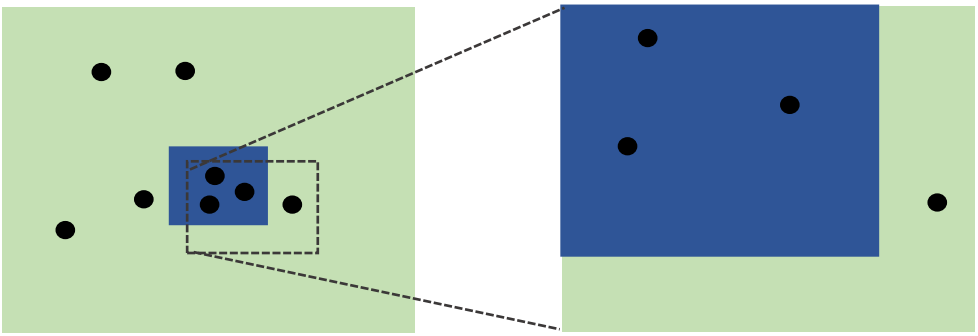


Figure 2: Visualization of the relativeness of language particularity. On the left, the languages within the blue zone seem to be very particular. The right image shows the effect of zooming in, where the languages within the green zone turn into rare languages.

To date, researchers have not treated this property of quantitative language indices. However, spatial analyses are often confronted with the change of sampling sets. Some examples are the result of zooming in or out of a region, and as the methodologies of this thesis are mostly based on spatial analyses, there is the need to discuss and quantify the effect of changing sample sets of language particularity.

2.4 Study Area

The analysis will be conducted in the Americas. The Americas include all three subcontinents of North, Central and South America and there are several reasons as to why this area represents an interesting region in terms of languages. Therefore, it is important to understand the history of the American migration.

The field of archeology has been able to date the settlement of the Americas. The analysis shows that migration to the Americas occurred across the Bering Strait, which was passable around 12'000 years before present. Archeologists have shown that before that time, the connection between Siberia and Alaska was covered by ice. With the retreat of the ice, the Bering Strait was passable from Siberia to Alaska. Due to the ongoing melting of the ice, the sea level rose until the Bering Strait was impassable again only some hundred years later. Consequently, there is a very clear time-window during which the Americas must have been populated. Archeologists agree that the Americas have been the last continent to be populated. For this reason, the Americas are also known as the 'New World' (Nettle 1999). Similar conclusions on the settlement of the New World have been reached in the field of genetics. Insights on genetics allow agreement on a single initial event of colonization of the Americas, which was followed by a few minor ones that are restricted to particular groups in North America. It is noteworthy to point out that the most extreme theory implies only one origin of the whole population in the New World (Wichmann et al. 2011).

Since the colonization happened across the Bering Strait, the origins of the migrants lie in northeastern Eurasia. This leads to the assumption that the first languages in the New World are based on that very same region (Nichols 1990). Due to the late colonization and the assumed small diversity of the communities, today's diversity of languages in the Americas would be expected to be very small. Surprisingly, a very high degree of language diversity can be observed within the Americas – compared to other continents. However, these languages and their history are still being discussed by different scientists.

Indeed, the first disagreements can be found in the definition of the term diversity. In the determination of the number of language families for the Americas alone, we can observe a discrepancy ranging from three to nineteen and up to 57 different families (Dahl 2008). In other studies, the language diversity in the Americas is illustrated by the number of linguistic stocks. Interestingly, 150 out of 250 linguistic global stocks are located in the New World (Nettle 1999). Other analyses show that South America represents the continent with the highest degree of language diversity. According to them, 60% of all isolated languages can be found in that region

(Dahl et al. 2011). Because of the high degree of diversification, which is indeed surprising, it is important to find reasons for such a development. Also, with regards to the meaning of high diversity, there is no agreement.

While some linguists take this unusual diversity as evidence for an earlier date of entry and try to falsify insights gained through archeology, others claim that there must be different components that have driven language development in the Americas (Nichols 1990, Nettle 1999).

One explanation lies in the non-linear development of language diversity. The model, visualized in Figure 3, suggests a high increase of language diversity at the beginning of the settlement of a new area. After around 20,000 years, language diversity decreases. Contrary to many other approaches, this theory would explain the high diversity of the Americas precisely because of the late colonization (Nettle 1999).

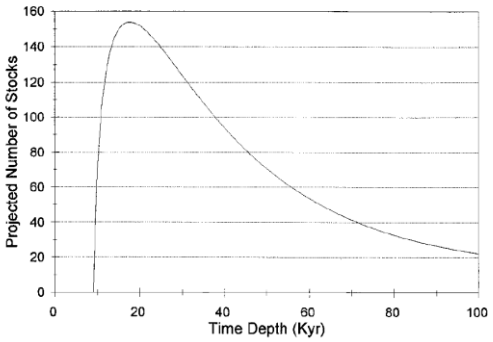


Figure 3: The development of language diversity over time (Nettle 1999).

Altogether it can be said that the Americas represent a very interesting study area based on the state of research within different areas. First, it is most likely that a small count of different communities had access to the Americas through a single place of entry within a small time-window. Thus, the Americas provide an isolated observation area. Secondly, the language particularity seems to be very high for the Americas, despite the fact that it is a big challenge to define language similarities. The high disagreement on this area amongst scientists implies that there is great potential for new insights and theories.

2.5 Data

The World Atlas of Language Structures (WALS) provides a dataset of 2679 languages around the world. 55 authors have gathered information about the family, central coordinates and structural properties of these languages. The genealogical information of all languages has been visualized in Figure 4, where each color represents one language family. Furthermore, the dataset provides

information about phonological, morphosyntactic and lexical properties, which are expressed in 192 different features. Each feature is described with different values. The number of values varies from two to twelve. For example, the feature 'Consonant Inventories' describes phonology by the use of five values from 'Small' to 'Large'. Unfortunately, not all features have been set for each language and the dataset represents a very sparse matrix. This means that each language contains a certain number of empty cells for undefined features and that each feature includes a specific number of empty cells, which do not explain a language (Dryer & Haspelmath 2013). Furthermore, features belonging to different types from A to G where A represents the base property and the ensuing letters some subcategories to it. In other words, the letters B to G autocorrelate with A (Dryer & Haspelmath 2013). The purpose of WALS is not primarily to compare languages, but to study particular features all over the world (Georgi 2010).

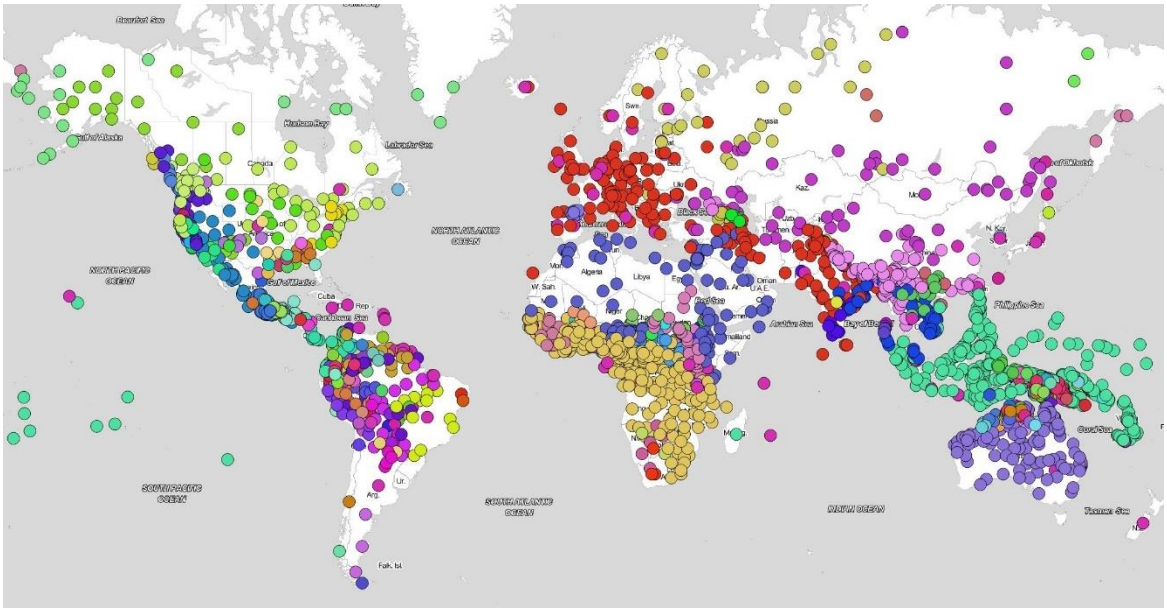


Figure 4: All languages of the WALS dataset, where each color represents one family.

2.5.1 Americas

The main analysis of this work focuses only on the subset of the Americas. The New World contains 654 languages that are categorized in 256 different families. As shown in Figure 5, the Americas are mainly represented through small language families.

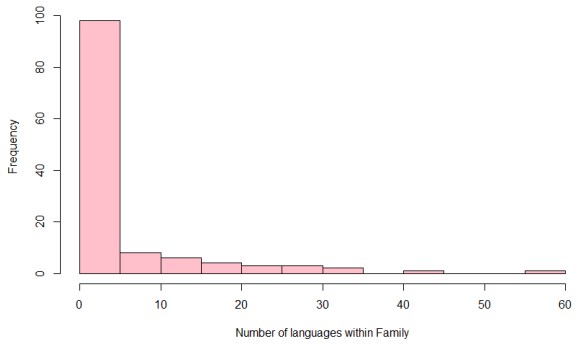


Figure 5: Number of languages within each family in the Americas

In total, there are 192 features that describe the languages of the Americas. Figure 7 shows the number of na's of these features. Considering that there are only 654 languages, there are features that only describe one or three languages. Figure 6 visualizes that the empty cells are distributed over the whole dataset. Most of the languages contain a high amount of empty cells and only few languages contain information of many different features. Overall, this means that the data set of the American languages contain 84% empty cells. Consequently, the data set requires a preprocessing where sparse features will be neglected as explained in chapter 5.1.

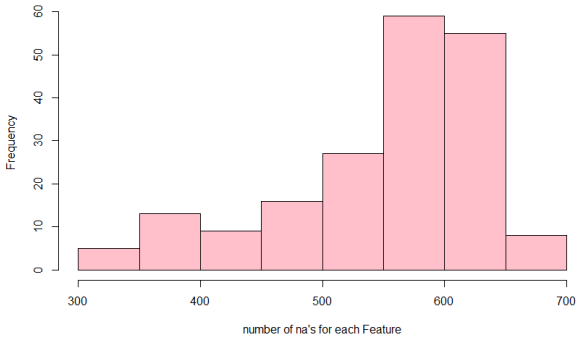


Figure 7: Summary of the number of empty cells for each feature in the Americas

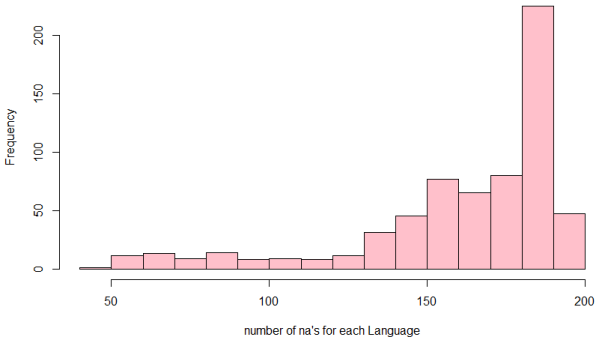


Figure 6: Summary of the number of empty cells for each language in the Americas

3 Geographical Dependencies

This chapter describes the scientific insights gained thanks to observing the impact of geography on language development. Languages are spoken all around the world, in different environments and societies. This spatial component turns languages into a geographical phenomenon. Hence, methods and theories used in the field of geography can be encountered. With regards to the disagreement on the unexpected language diversity in the Americas, this thesis has already discussed the spatial influence on a very large scale. The development of languages in different places around the world might differ or be influenced by different factors. This phenomenon is known as *geographical* or *spatial dependencies* and includes factors of the environment such as land surface, climate or weather condition. To determine spatial dependencies, geographic information science (GIS) provides different approaches and methods. With the help of quantitative and spatial analyses, several correlations between languages and geographical elements have been shown. The next section presents the most important insights in terms of languages and how they are influenced through geographical dependencies and geographical isolation.

3.1 Languages and geographical Isolation

One of the most common theory in geography, known as *Tobler's first law*, describes: "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). This law can also be applied to languages. As a result, geographically close languages are more smiliar to each other than languages that are geographically distant. Holman et al. (2007) found a correlation between metric distances and language similarities over different continents, which determines this assumption. In the field of genetics, this phenomenon is known as '*isolated by distance*' (IBD). In other words, the greater the metric distance between two observations, the greater their isolation and the fewer similarities can be found between them. Therefore, this illustrates that metric distance describes a simple concept of isolation.

In terms of societies, isolation implies a small probability of contact to others. Thus, the development of languages in isolated societies is mainly internally-driven and the languages are assumed to be more particular. The most acclaimed linguistic analysis in IBD has focused on dialects instead of languages and the scientific insights for IBD concerning languages are only sparse (Wichmann et al. 2011). One study verifies the concept of IBD for languages based on the method of clustering. More specifically, languages have been clustered through spatial subsets and compared to other randomly sampled language clusters. Regardless of whether the size of the

sample and the geographic region has been taken into account, spatial subsets have always shown more similarities in language traits than random subsets (Wichmann et al. 2011). This outcome indicates the correctness of Tobler's law in terms of languages. More detailed analyses show that the impact of distance on language particularity decreases by increasing distance between languages. Furthermore, the correlation coefficient regarding distance is higher for related languages than non-related languages (Holman et al. 2007).

As already pointed out, metric distance is a factor of isolation. In this manner, also the term of *accessibility* will be used, whereas accessibility is mainly applied for *low isolation*. Within this thesis, isolation is always understood as isolation in terms of human movement, which is highly influenced by other factors than metric distance. Some examples are topography, environment, climate, culture and religion. The process of internally-driven language change described in section 2.2.1 dominated the development of isolated languages.

In the matter of human movement, the opposite of isolation can be defined as *human contact*. Thus, a high degree of isolation implies a low degree of human contact and the other way around. Since the origin of languages lies in the wish of humans to communicate with each other, the encounter of different societies has highly influenced language development (Heggarty 2015). The movement of societies allows the contact of different communities and leads to contact-induced language change. Since human movement is highly influenced by the environment, some routes are more probable to be traveled than others and thus increase probability for contact between societies. Therefore, the more detailed the human movement can be reconstructed, the more insights can be gained with regards to language behavior (Rogers et al. 1990).

Overall, there are only a few studies that have combined languages and geographic factors other than distance. Analyzes have shown a correlation between genealogical language diversity and the amount of rainfall (Nettle 1999). Other studies have shown that coastlines or tropical areas have a tendency to high language particularity or that the latitude positively affects diversity (Nichols 2003). In these examples, heavy rainfall, as well as tropical and mountainous areas, cannot be seen as the direct cause for language development. Rather, they are more likely represent indicators for isolation.

Instead of investigating the influence of geographical isolation on languages, the impact of human contact on language diversity can be examined. Rogers et al. (1990) for example proposes that today's languages can be explained based on historical environment datasets. During the glacial Wisconsinan (18'888-14'000 BP), the only ice-free corridor from Beringia to the south of America

was located on the western coast of North America. Assuming an entry point to the Americas over the Bering Strait, this region defines the major route and the area with the most human contact (Rogers et al. 1990). Languages that have been isolated by ice have developed differently than languages that have been in contact with others. Such zones of contacts defined by several environmental and cultural components are also called biogeographic zones. Within these zones, languages are prone to adopt features of other languages (Rogers et al. 1990). In other words, the languages of societies that are likely to get in contact with others are less particular than languages that have been isolated.

Other studies have shown that areas with constant societies that stay there over a long time have greater language diversification than areas that have experienced more change in occupation. Increasing change in occupation indicates more contact between different societies and consequently generates a zone of contact (Gruhn 1988).

All these studies lead to the very same conclusions. First, the geography and environment have a major influence on the change of languages. Second, the influence of the environment is a result of the degree of isolation and probability of contact. Until now, the exact influence and processes of language change due to contact has not been spelled out yet. However, many linguists acknowledge the influence of isolation on languages (Nichols 1990).

Isolation and accessibility, respectively, are very vague concepts. There is neither a common term to describe nor a common method to measure them. While some call it the level of isolation or connectivity, others speak of biogeographic zones (Murrieta-Flores 2012, Rogers et al. 1990). This thesis uses the terms of accessibility and isolation whereas low accessibility means high isolation and the other way around. This accessibility has been described and conceptualized in many ways. Some examples include metric distance, altitude, conflicts of societies or historical movement. Nevertheless, all these factors of accessibility have shown an influence on language particularity. Given these findings, this thesis assumes that the level of accessibility due to different environmental factors highly influences language particularity. Thereby, zones of higher accessibility are more likely to present a zone of contact, where languages approximate each other. Zones of low accessibility result in more particular languages. The hypotheses will be discussed in detail in chapter 4.1.

To test these hypotheses, a quantitative model will be developed that allows computing the value of accessibility around societies and their languages. In order to achieve that, methods of geographic information systems and theories of archeology will be considered predominantly.

4 Lack of research

Scientific research has shown that the topic of language diversity, similarity or rarity and the use of quantified measurements has recently gained a growing importance in historical linguistics. Nevertheless, there is a clear lack of clarification, verification and discussion on the methods and indices of such particularity measurements. In most of the literature, the methods for defining the similarities of languages are not explained at all. Mostly, it is not even clear whether the particularity is meant genealogically or typologically. Only a few linguists have discussed quantified particularity indices in detail. These findings have been used by many without further investigation. Moreover, the important issue of scaling impacts in the field of geography has not been questioned at all. Since a language can be defined as similar or not similar compared to another language, language particularity is a relative measurement of the sample set. Spatial distribution is a crucial factor, including the question, for example, on which continent a language is spoken. Therefore, a discussion about scaling-issues and sampling dependencies is inevitable. As a result, the research question number 1 has been formulated.

Many linguists have shown the importance of the degree of contact and isolation for language development. Even though there is agreement in qualitative theories, only a few have encountered quantitative analysis to verify such common assumptions. The few analyses that have been carried out have mostly considered simple factors of isolation such as distance, altitude or the amount of rainfall. Due to the specific study area and the lack of transparency, some of these outcomes first need to be verified in research question number 2. Furthermore, there is shortage of appropriate concepts and methods to describe contact of societies and their impact on languages. In order to address this issue, methods from the field of archeology are combined with GIS and linguistics in research question 3. The aim of RQ number 4 is to acquire knowledge to the change of languages due to the environment. In order to do so, the insights of qualitative languages indices and quantitative indices to describe isolation accessibility are combined respectively.

4.1 Research Questions

RQ1. How do scaling-issues affect quantitative language particularity?

The particularity of a language is highly dependent on the sample set of comparison. This leads to the assumption that the particularity of a language compared to all languages of the world is different to the particularity of languages that are merely compared to its nearest neighbors. In other words, language particularity is a highly relative measurement.

Hypothesis 1: Language particularity is highly dependent on the scaling-levels.

RQ2: How do simple factors of isolation explain language particularity in the Americas?

Different studies have shown that simple factors of isolation can already describe language particularity to a notable degree. It is not clear for all studies which measurement of language particularity was used for this correlation to be found. Due to this and because of the specific study area of the Americas, some of these outcomes are going to be verified in a first step.

Hypothesis 2: Metric distance and altitude are simple factors for isolation and can explain language particularity in the Americas.

RQ3: What is an appropriate method to model accessibility in terms of language particularity?

The scientific research revealed a lack of quantitative analysis in linguistics in terms of human movement. Also, it was shown that the combination of archeological models and methods from geographic information science (GIS) offers a lot of possibilities. The research question of this thesis is about an appropriate concept to model accessibility and its implementation.

Hypothesis 3: Archeology and GIS provide appropriate methods to model accessibility in terms of language particularity.

RQ4: What is the impact of accessibility to language particularity?

The main goal of this master thesis is to get new insights on the impact of accessibility to language particularity. Based on the results of the previous RQ 3, it will be feasible to determine new perceptions and theories on language behavior due to the environment. Thereby, a discussion about the parameter of the model is crucial.

Hypothesis 4: The level of accessibility due to environmental factors highly influences language particularity in a negative correlation.

Chapter 5 presents the methodology to discuss the research questions. It will be shown that the field of GIS provides methods and concepts that can be applied to reach new insights in the field of linguistics.

5 Methodology

The previous chapters gave an overview of the state of art and the lack of research. This chapter determines and explains the methods that will be used to answer the research questions. Therefore, different methodologies must be determined to define all variables listed in Figure 8. First, the approach to define *genealogical* and *typological particularity* are explained. To tackle RQ 1, different levels of typological particularity are described in section 5.1.3. The four variables to model the degree of isolation are presented in section 5.2. Whereas *IBD* and the *altitude* present rather simple approaches, the more detailed method to create *reachability Polygons* is treated in *Accumulative Costs*. Based on this, the variable of the *Area of Contact* is explained.

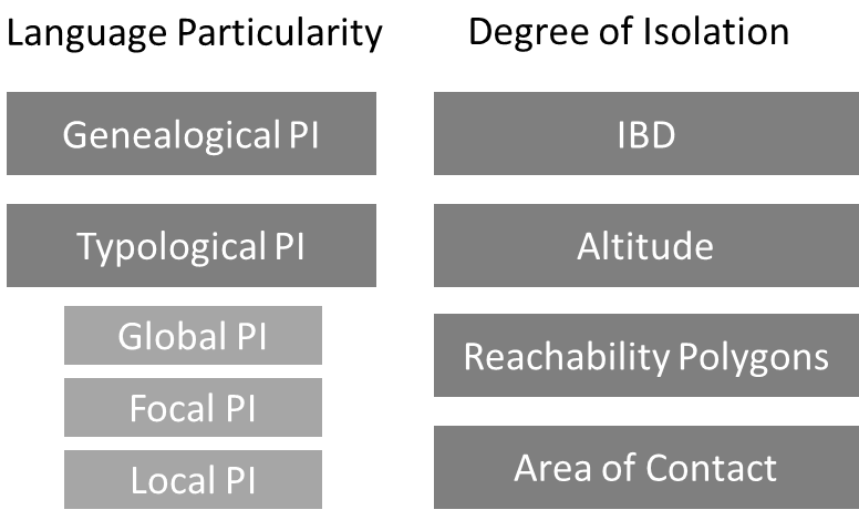


Figure 8: Variables to define language particularity, resp. the degree of isolation to tackle research questions 1-4.

5.1 Language Particularity

The two different approaches that define language particularity were introduced in Chapter 2.3. Genealogical particularity, as well as typological particularity are both the result of comparative statistics. The following section describes the methods and formula to create these indices in detail. Comparative statistics are always dependent on the objects that are compared with each other. By changing the sample set, the results are likely to vary. The approach to determine the impact of the different sample sets on language particularity will be described in section 5.1.3.

5.1.1 Genealogical particularity

The first and simplest way to determine language particularity is based on the genealogical descent of each language. This approach assumes that languages within the same family are more similar than languages from different families as described in chapter 2.3.1. This means that languages of large families are less particular than languages of small families. As a result, particularity can be computed through simple probability calculations. The probability measures the degree of likelihood to which a language belongs to a specific family. The probability *P(x)* for each family can be computed by:

$$1) \quad P(x) = \frac{lang\,(\mathrm{fx})}{\mathrm{tot\,(lang\,)}},$$

with the number of languages lang in family *fx* and the total number of languages *tot(lang).* With regards to the goal of this analysis, particularity is more important than probability. Importantly, particularity *Part(x)* that is the counterpart of the particularity and can simply be calculated by:

$$2) \quad Part(x) = 1 - P(x).$$

The termination of genealogical PI is used from now on to describe the index of genealogical particularity.

5.1.2 Typological Particularity

The quantitative analysis of language typology to describe language particularity has not been used often. Its principle is similar to the index defined through the family membership. The main difference is that the particularity is going to be described as the mean value of all present features. The index measures the likelihood of the features of a language on average over a certain region.

The first challenge occurs due to the structure of the dataset. The WALS dataset is represented by a sparse matrix; accordingly, there are a lot of *empty* cells that are also called non-available information (na). In average, the features only describe 20% of all languages. Thus, 80% of the features contain na. The distribution of the na's is not regular. In fact, there are some features that clearly contain more information. Additionally, there are many languages only described by a few features and there are only very few languages described by a high number of features.
In quantitative comparison, languages that only contain very little information, as well as features that have a lot of empty cells, are problematic. The question of the maximum number of allowed non-available information for quantitative analyses has been treated by very few different scientists

only (Cysouw 2015, Daumé 2009). Most of the methods have been defined depending on the global dataset. As the dataset of the Americas is much smaller, not all of them are appropriate for this study area. The most suitable method for this analysis has been suggested by Daumé (2009). This method removes all features that appear in a quarter of the languages at most. As explained earlier, all features are categorized from A to G. To prevent autocorrelation, all features that are not of type A are eliminated. In a next step, languages that are only poorly described are removed. The method of Daumé (2009) suggests including only languages that are described by at least 11 features. For the Americas, this results in a dataset of 342 languages and 37 features.

In comparison to the genealogical PI, the typological particularity summarizes the probability of multiple features. Thus, it represents a mean value of probabilities of all features. The detailed formula is shown below and presents formula 3) to 7). An additional difficulty is the variation of the number of values that can be used to describe a feature. For this reason, the average probability had to be normalized by this number of values as shown in formula 7). The detailed procedure will be explained below with help of the formula 4) to 8).

The typological particularity can be computed based on the following information for each language. Language A is described through n features. For all n features, information on how often each feature is used to describe a language (*Count_Feature)* is necessary. Furthermore, the number of categories that is used to describe each feature is required *(Count_Values)* as well as how often each of this category has been used in total *(Count_of_Values).* The resulting *wide table* illustrates these necessary values. Table 1 lists all possible combinations of Feature, Value and Name from the original dataset. The language *Émérillon* is amongst others described through the feature *Coding of Nominal Plurality* and the value *8 Plural clitic*. This feature describes 243 other languages and is defined through seven different values. It follows that the value 8 *Plural clitics* explains 14 languages. Thus, the probability of this feature value *rprob* can be computed like so:

$$3) \qquad rprob = \frac{Count\_of\_Value}{Count\_Feature}.$$

The particularity *part* is reached by:

$$4) \qquad part = 1 - rprob.$$

In this specific example, the particularity is reflected in the small value of *rprob* that results in a high value of *part*.

Table 1: Extract of the wide table where all variables to compute TP are summarized. Count_Values: number of values for each Feature, Count_Feature: number of languages each Feature describes, Count_of_Values: Frequency of each value - Feature combination

Feature	Value	Name	Count_Values	Count_Feature	Count_of_Value	rprob
Coding of Nominal Plurality	Plural clitic	Émérillon	7	243	14	0.0576
Position of Case Affixes	Postpositional clitics	Émérillon	6	247	33	0.1336
Position of Pronominal Possessive Affixes	Possessive prefixes	Émérillon	4	254	159	0.626
Poosition of Tense Aspect Affixes	Tense-Aspect suffixes	Émérillon	4	273	194	0.710

Because of the unequally distributed number of values across the features, the probability of features with few values is automatically higher. To avoid this bias, the probability and particularity must be computed respectively through a variable *inv* by the inverse of the number of values of a feature *Count_value*:

$$5) \qquad inv = \frac{1}{Count\_Value},$$

With the help of this variable, the weight *w* can be defined:

$$6) \qquad w = \frac{inverseCountValue}{\sum(inverseCountValue)},$$

in so doing, the particularity can finally be normalized *partNorm* with the help of equations 3 to 6:

$$7) \qquad partNorm = \sum(part * w).$$

The described method to define typological particularity has been used most recently in linguistic studies and seems to be the most suitable for this analysis. However, the results in chapter 6 will show that this method leads to a lot of unexpected problems. Thus, an alternative method has been established. The method refers to Cysouw (2016) that provides a package called *qlcMatrix*. The function *Similarity-measures between objects* (*sim.obs)* has been developed to define similarity of observations that are described in nominal data. Coincidentally, the original data used by Cysouw is the same as used in this thesis, namely the WALS data set. Therefore, the method has explicitly been developed to handle sparse matrices.

In here, the principles of the function are described. The detailed code is provided as open source (Cysouw 2018). In difference to the average mean described above, this method compares each pair of languages. Thereby, only the existing properties are compared and na's are ignored. This is called

a *complete case analysis*. Within a contingency table, the accordance of all variables of the language pairs is summarized. In other words, the contingency table counts how often the properties of languages are identical. After a normalization, these values are used as the observation values *O.* Based on the probability of the marginal distribution, an expected value *E* can be computed such as a pointwise mutual information *pmi* value can be computed:

$$8) \qquad pmi = \log(\frac{O}{E}).$$

The pmi describes the discrepancy of the observed to the expected value. The less commonalities we have, the smaller the pmi and, consequently, the larger the particularity is. In this thesis, this value will be called *typological PI*.

5.1.3 Relativeness

The process to define particularity as shown above confirms that each language particularity is a relative measurement compared to other languages. This thesis tries to determine the meaning and impacts of this *relativeness*. Indeed, this will be shown in the example of the typological PI.

The typological PI will be defined based on different spatial subsets as visualized in

Figure 9. The first one uses the whole study area of the Americas and all American languages to define each language particularity. This will be called the *global PI*. The second one determines language particularity for each subcontinent, the *focal PI*. This means that all languages of South America are compared with the languages of South America and all northern languages with the subset of the north and central American languages. The last one is called the *local PI*. The local PI defines the particularity of a language based on its nearest neighbors; for example, all languages within 1000 km.

Some first data analyses in chapter 2.5 show the high percentage of non-available information within the language data set. Especially for the local PI, this can be problematic. The fewer languages are compared, the higher the probability is that we do not have information about the same features. Namely, if language A gets classified using feature a, b, c but language B using feature d, e and f, they cannot be compared with each other. To avoid this, the same data reduction is necessary for each neighborhood, according to Daumé (2008). All features that explain less than ¼ of the languages must be evaluated and neglected.

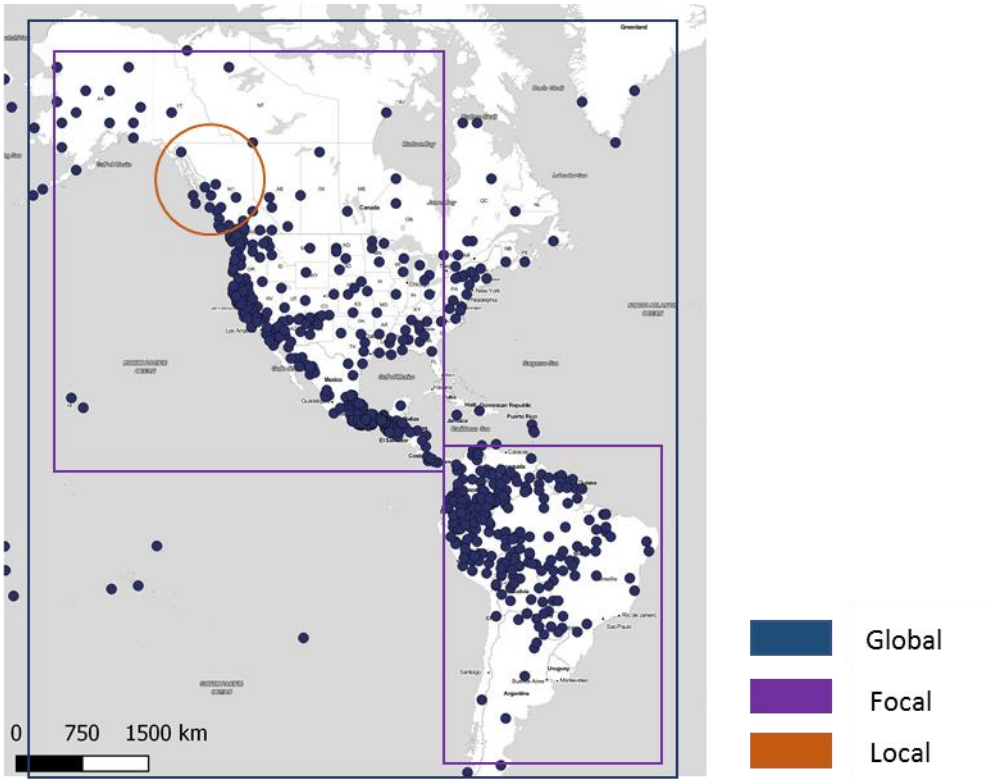


Figure 9: The language particularity for each language will be computed based on three different sampling sets. The global sampling set includes all American language points. The focal defines particularity in comparison to all languages of the same subcontinent and the local PI only considers languages within a neighborhood of 1000 km.

5.2 Geographical Isolation

To analyze the exact influence of accessibility on language particularity, one needs a quantitative model of accessibility. Most of the methods to model human movement, which spans across landscapes, have been developed in the field of archeology. In section 3.1 it was shown that the concept of accessibility is poorly defined and, consequently, can only be measured inadequately. As different situations and purposes demand different approaches to model accessibility, a global definition cannot be given (Geurs & van Wee 2004). In order to elaborate research question number 3, these concepts will be combined to find an optimal strategy to implement accessibility in terms of human movements. The simplest concept of isolation is based on metric distances. The method to model *isolation by distance* is explained in section 5.2.1. Least-cost-path analyses (LCP) allow the inclusion of further information. It can be concluded that the use of topography presents the most established component in terms of human movement. This approach will be adapted for this analysis to model accessibility. The core elements of the LCP's, that is, mainly the cost surface, are discussed in section 5.2.2.

5.2.1 Simple Indicators of Isolation

Isolated by Distance

The simplest way to model accessibility is the use of metric distances as the only component. This means that the greater the metric distance between two observation points, the greater their isolation and the fewer similarities can be found between them. Based on this assumption, the model of 'Isolated by Distance' (IBD) has evolved within the field of Genetics. Additionally, IBD models have been used for language analysis and it was shown that in many regions of the world language similarity can be explained though the metric distance between two languages (Holman et al. 2007).

The main components of IBD analysis are metric distances and thus distance matrices. By the use of distance matrices, the enhanced similarity of language properties for close languages can be determined. This is also known as autocorrelation. In association with autocorrelation, the term of *neighborhood* is often used. A neighborhood is mostly defined by the neighbors within a specific radius or by the number of nearest neighbors (Sokal & Wartenberg 1983). Considering IBD, the assumption may be formulated that the smaller the count of neighbors within the neighborhood, the more isolated and the more particular the language is.

For each language point, the number of neighbors within a radius r can be determined as visualized in Figure 10, and compared to its particularity. Alternatively, a specific number of neighbors can be defined and the mean distance to them can be computed. Both methods are assumed to result in a positive correlation. In Figure 10, this means that the language on the left is supposed to be less particular as the language on the right, because of the higher number of neighbors.

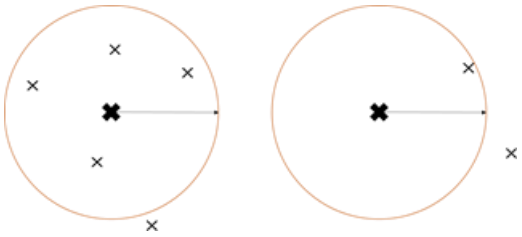


Figure 10: Amount of languages within a specified radius

Without a doubt, metric distance is fundamental in human movement, as humans are more likely to get in contact with less distant communities. IBD and autocorrelation provide simple analyses to gain first insights about the spatial pattern and the relation to a society's neighbors. However, there are a lot more components that influence human movement in the environment than the simple metric distance.

Altitude

As shown in section 3.1, *altitude* of each language point seems to reflect isolation. Specifically, altitude describes the height above the sea level. In essence, high altitude is likely to represent a mountainous or rough area and low altitude indicates flat and well accessible regions (Nichols 2013). To address RQ2, Figure 11 visualizes the correlation analyses to be determined between the following variables. Particularity has been defined as genealogical, global, focal and local PI. The variable of the degree of Isolation and reachability is described in form of IBD respectively and the altitude so far. Chapter 6.3 will determine the correlation between these variables and will show that the variable of local PI is not suitable in terms of the correlation analysis.

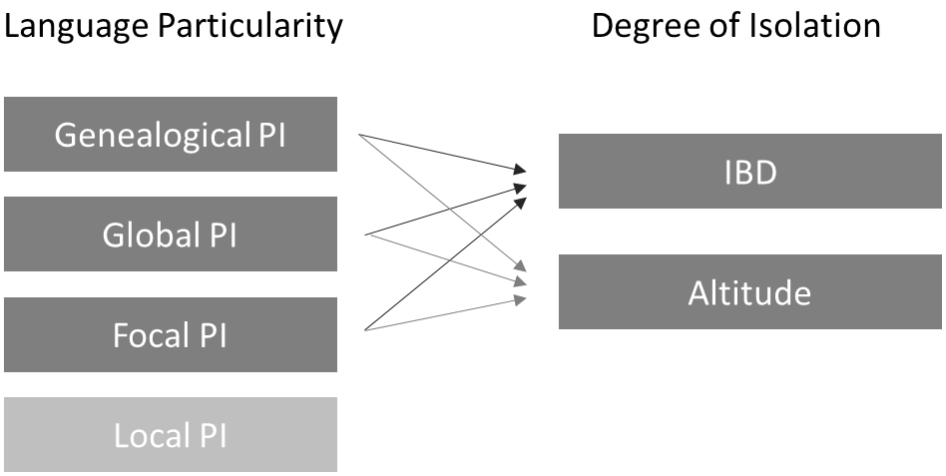


Figure 11: Visualization of the regression analyses that will be conducted based on the created variables to define language particularity and the degree of isolation.

5.2.2 Accumulative Costs

Overview

Besides IBD analyses, there is the need for more sophisticated methods to include further information. This section describes the methodology for defining isolation based on accumulative cost surfaces. Therefore, methods and concepts of LCP's as well as hydrological network analysis will be combined. First, the basic concept of the methodology is introduced. Second, the exact and detailed implementation is described in steps one to 7.

Besides distance, the most common component to model human movement is topography. Topography allows for definitions of slope and aspects or roughness of an area, which are all indicators for isolation. A very steep area, for instance, is harder to travel through and to reach than a flat region (Howey 2007). Contrary to common GIS approaches, slope is not only considered as

absolute values, but rather, as how people experience the terrain. This experience of the terrain can be represented as a cost surface. To illustrate, a cost surface is represented in form of a raster layer where each cell refers to a specific cost. In fact, these costs describe the difficulty of moving in between the cells (White & Barber 2012). Significantly, cost surfaces have also a very high importance in LCP analyses. LCPs are a common method for route analyses in GIS. As the name indicates, the effort to travel is expressed as *costs*. In most of the LCPs, these costs are defined based on the elevation model where each cell of a raster is assigned by a cost due to the properties of the elevation model. The layer of the costs is then called *cost surface*. LCPs can determine the best route between any two points on this surface called the *origin* and *destination point* (Howey 2011). The insights of LCP studies are used to define cost surfaces. If the cost is depending on more than only one factor, for example, elevation and land surface, it is known as multiple criteria cost surface (Howey 2007). As pointed out in earlier chapters, the use of further influences has been evaluated, but rarely used in terms of GIS analyses. Due to the fact that the cost surface represents the core element of the analysis, the determination of these costs is crucial. The exact definition of all costs will be explained below in step number 2.

LCPs evaluate the route with the least costs between two points. However, in term of historical human movement, the assumption of a known point of destination is unrealistic. Only a few scientists have defined models that allow the definition of accessibility without an origin and destination point. This thesis adopts the approach of accumulative cost surfaces. These cost surfaces have originally been developed to model hydrological networks. The basics of hydrogeological models also lie to cost surfaces. For each cell, an accumulative value is defined based on the number of cells that flow into this cell (Llobera et al. 2011). So far, this means that the costs of moving from point A to point B can be defined as

$$9) \qquad cost_{AB}(DHM) = cost_{A}(DHM) + cost_{B}(DHM),$$

where $cost_{AB}$ (DHM) defines the costs of moving from one cell into another on a terrain that is generated due to the costs of cell A, based on the elevation as $cost_{A}(DHM)$ and the costs of cell B of the elevation $cost_{B}(DHM)$. An example that illustrates such accumulative costs has been visualized in Figure 12. Starting from one point, all costs that are traveled through must be accumulated. In the end, the accumulated cost defines the total effort that has been applied to travel this route.

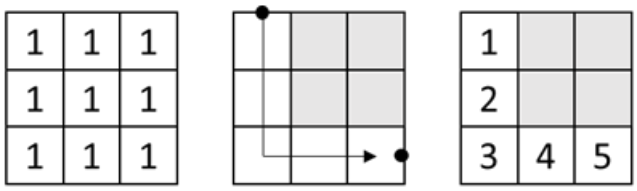


Figure 12: Cost Surface and the resulting accumulative costs

While hydrological network analyses and LCP aim to find the path of least resistance, this analysis is about finding the resistance of the region in general. Thus, instead of only accumulating the costs for a single route, all accumulative costs can be computed in each direction and for each cell, as visualized in Figure 13.

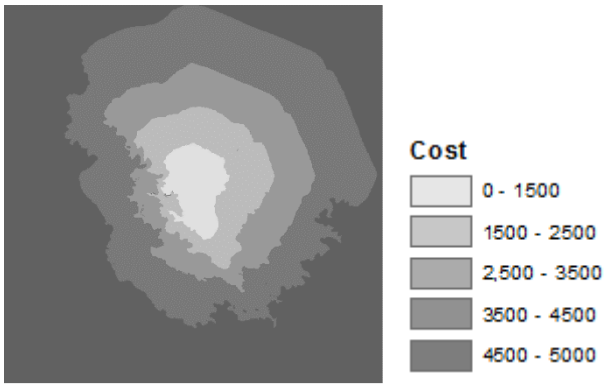


Figure 13: Computed accumulative costs from one point in each direction

Howey (2007) points out that the costs of traveling through a landscape are not only influenced by topography. Environmental features, such as water bodies, vegetation, natural barriers or cliffs, as well as cultural resistance, are highly influencing the route choice in terms of human movements. The way water bodies are discussed varies in different studies due to the assignment of costs. Some claim that they are natural barriers that cannot be crossed by pedestrian movement, but only by canoeing (White & Barber 2012). The oldest canoe paddle in the Americas has been traced back to 1300 BC (McKillop 2005). Consequently, when modeling human movement before this time, waterways should be treated as insurmountable barriers, which would otherwise pass as easily passable surfaces (Howey 2007). Independently of the fact whether waterbodies have been crossed or not, especially traveling along rivers has been vital in human movement. Therefore, water bodies also influence the possibility of contact and development of languages (Ranacher et. al 2017). As

Rogers et al. (1990), this analysis includes historical surface as a second component for historical human movement.

Thus, the accumulative costs based on the DHM and based on the historical land surface are understood as the level of accessibility. The higher the costs are, the less accessible the according region is. A specific threshold of the costs is assumed to represent the maximum energy humans can afford to travel. Using this specific energy, a larger area can be reached in each direction within a flat area than within a steep one. For this thesis, the area of the maximum threshold is called *reachability polygon*. Thus, the area of the reachability polygon A within a flat terrain is larger than the one within a steep area of language B. Consequently, the center of the region A is better accessible and less isolated as the center of region B, respectively. Thus, the area of the accessibility polygon represents isolation.

In summary, the accumulative cost polygons avoid the unrealistic assumption of common movement analyses in which people in historical human movements have been traveling to a known destination. The cost surface is flexible to add different layers and different weights as shown in step number 4. The accumulative costs allow the definition of a general degree of isolation without pairwise comparison. Nevertheless, the size of the threshold cannot be verified through scientific research. It is noteworthy that with a too big of a threshold the analysis loses the advantage of its lower computational complexity. The method allows for a general definition of isolation for each point.

In the next section, the described method is explained in detail. For this purpose, we are taking a closer look at the datasets and the exact parameter space. In step 1, all information about data and preprocessing is given. The assignment of all costs is explained in step 2. Based on the cost surface, a so-called transition layer is computed as shown in step 3. Moreover, in step 4, the parameter of cost transformation and layer weighting is put across. Step 5 includes the necessary step of geocorrection to provide a realistic model. The resulting accumulative costs and the definition of a threshold to determine reachability polygons are explained in step 6 and 7.

1. Data and Preprocessing

Elevation Model

The elevation models of North and South America have been provided by the USGS GTOPO30 (USGS 2015). The resolution of the grid is defined through 30 arc seconds that correspond to an average cell length of 700 meters and are visualized in Figure 14.

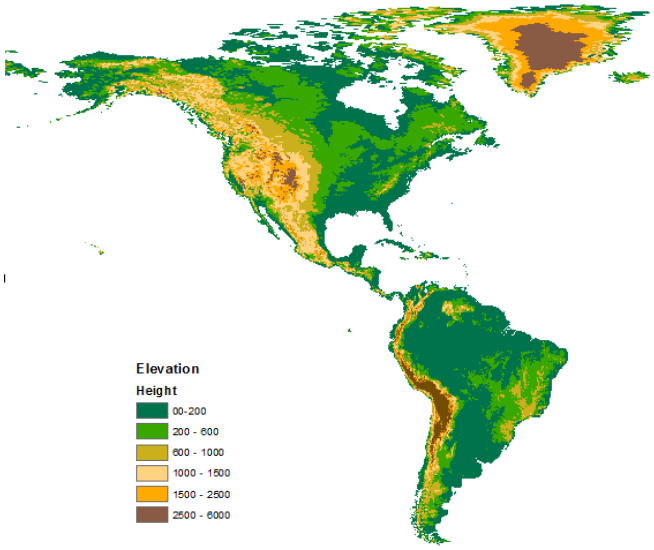


Figure 14: Elevation model of North and South America

Historical Land surface

Language change has a time depth of many thousands of years (Heggarty 2015). It was shown that environmental factors such as, for instance, land surface have a high influence on human movements and language development. Thus, the languages of today have not been shaped by the actual but by the historical land surface (Rogers et al. 1990). For most realistic analyses, the language data could also be provided for historical spatial distribution. However, the lack of such historical linguistic datasets presents insurmountable limitations. Therefore, this analysis explains today's languages through historical land surface.

Archeologists have modeled the world's land surface from around 18'000 years before present and have called it *Intarch*. This time corresponds to the last glacial maximum. The migration to the Americas took place around 12'000 years before present. Thus, the data set of *Intarch* has been evaluated as the most suitable model for this time span. As visualized in Figure 15, the Americas are categorized into 21, broad-scaled different land covers (Ray & Adams 2001).

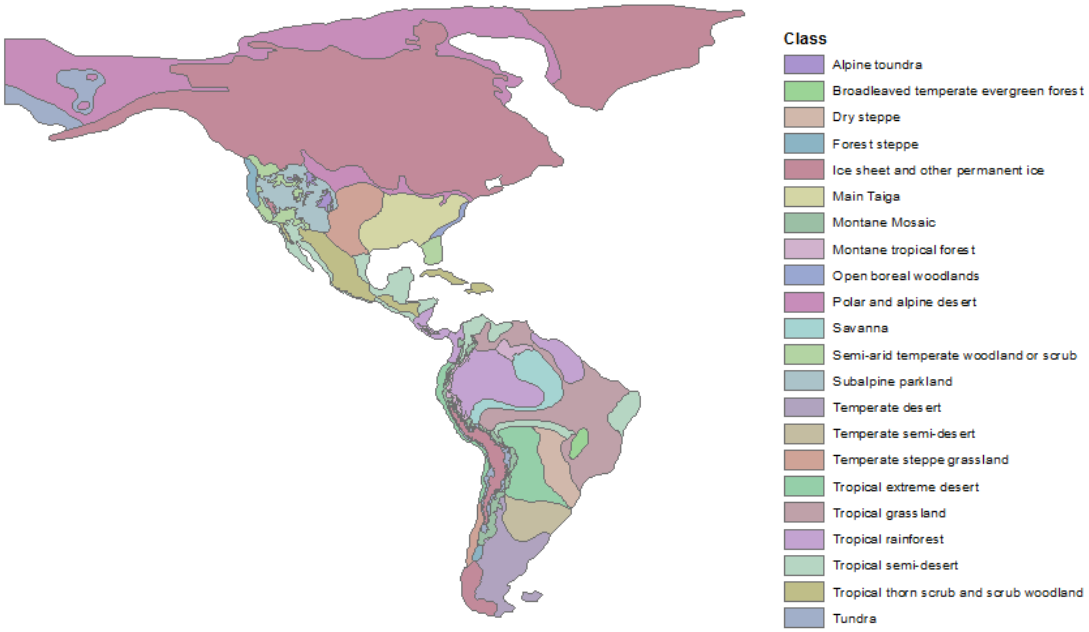


Figure 15: Modeled historical land surfaces of the Americas 12'000 before present (Ray & Adams 2001).

2. Cost Definition

As already pointed out, the definition of the costs is crucial for accumulative cost analyses. However, there is not a general theory to define these costs. Also, there is a large difference in the state of research to define costs depending on the data layer. While the moving costs for the topography are well established, a lot of uncertainties exist for moving on different land surfaces.

Elevation

In most of the accessibility analyses, elevation describes the main component of human movement. In fact, simple analyses have shown the importance of the elevation only through the value of the altitude. To describe movement within topography, Tobler (1993) introduced the so-called hiking function:

$$10)\ cost(DHM) = 6 * \exp(-3.5 * abs\left(\frac{dh}{dx} + 0.05\right)),$$

in which the variable *cost(DHM)* describes the velocity of moving and where *exp* computes the exponential of the constant variable *-3.5* and the absolute values *abs* of the slope that is defined through the difference of height *dh* and the length of the cells *dx,* that defines the slope, added to the constant variable of *0.05.* It estimates the speed in which humans can travel within different degrees of slopes.

As shown in Figure 16, the relation between speed and slope is not linear in formula 10. The steeper the slope, the flatter the curve of speed. For example, within a slope of zero degree, the speed of movement is set to 5.03 km/h and for a slope of 15° to 1.07 km/h (Van Etten 2017).

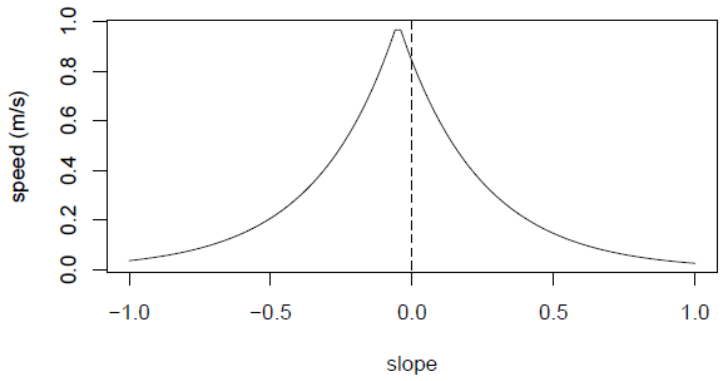


Figure 16: Value distribution of Tobler's Hiking Function

Historical Land Surface

As discussed in previous chapters, the use of historical land surfaces is crucial for pedestrian movements. This means that there are additional costs when moving from point A to point B. The costs of moving within the historical surface can be defined similarly to the elevation layer:

$$11)\ cost_{AB}(Surf) = cost_{A}(Surf) + cost_{B}(Surf),$$

where, again, the accumulative costs from A to B on the historical surface $cost_{AB}(Surf)$ are the accumulation of the costs of cell A $cost_{A}(Surf)$ and cell *B* $cost_{B}(Surf)$. Thus, the historical surface must be categorized into costs. Unfortunately, the rather new approach of quantitative movement models, the high dependency to the study area and the resolution of the input data complicate the finding of a common denominator for cost assignment.

First qualitative investigations on human movement on different land surfaces have been conducted in the 70s. Different terrain coefficients have been defined for different terrain. The terrain coefficient in Table 3 represents the level of resistance, instead of costs (Soule and Goldman 1972). White and Barber (2012) have accomplished the categories and defined actual costs for similar classes shown in Table 2.

Table 3: Land cover classes and their terrain coefficient (Soule & Goldman 1972)

coefficient system used by the edge cost generator discussed below (Soule and Goldman, 1972).

IGBP Class	Description	Terrain coefficient
0	Water	1.8
1	Evergreen Needleleaf forest	1.5
2	Evergreen Broadleaf forest	1.5
3	Deciduous Needleleaf forest	1.5
4	Deciduous Broadleaf forest	1.5
5	Mixed forest	1.5
6	Closed shrublands	1.2
7	Open shrublands	1.2
8	Woody savannas	1.2
9	Savannas	1.2
10	Grasslands	1.0
11	Permanent wetlands	1.8
12	Croplands	1.2
13	Urban and built-up	1.0
14	Cropland/Natural vegetation mosaic	1.2
15	Snow and ice	1.5
16	Barren or sparsely vegetated	1.0
254	Unclassified	0.0
255	Fill Value	0.0

Table 2: Grouped land cover and their cost assignment by White and Barber (2012)

Grouped land cover	Final assigned cost value	Reasoning
Major River	5	Canoe travel occurred up major waterways
Lake	65	Avoid "puddle-jumping"[a]
Great Lakes	10	Canoe travel occurred along shoreline
Forested Wetland	25	Part of waterways – not as easy as unobstructed river
Non-Forested Wetland	70	Impassable wetlands with canoe, wet walking
Forested	40	Vegetation but openings
Non-Forested	30	Easy to walk/see through
Sparsely Vegetated	60	Includes sand and rocks, want to avoid scrambling
Natural Disturbance	100	Random, unpredictable occurrences

Neither of the studies can be translated one-to-one with regards to the data of the historical land surface. However, the combination of both allows estimating the costs for the historical land surface of the Americas as visualized in Table 4. The highest costs of transition for land surfaces have been set to *Ice sheet and other permanent ice*. The next costly surfaces are all extreme deserts, followed by different kinds of forests. *Savannas*, *Tundras* and *Steppes* are the next category. There is one disagreement in the literature concerning the travel through water. Because this work assumes a likely travel along water bodies, the lowest costs have been assigned to the class of *Water*.

Table 4: Cost assignment of the historical land surface of the Americas based on Soule and Goldman (1972) and White and Barber (2012)

Class	Cost
Tropical rainforest	40
Monsoon or dry forest	40
Tropical woodland	40
Tropical thorn scrub and	40
Tropical semi-desert	15
Tropical grassland	15
Tropical extreme desert	60
Savanna	25
Broadleaved temperate	40
Montane tropical forest	40
Open boreal woodlands	25
Semi-arid temperate	25
Tundra	25
Steppe-tundra	15
Polar and alpine desert	60
Temperate desert	25
Temperate semi-desert	25
Forest steppe	25
Montane Mosaic	15
Alpine toundra	15
Subalpine parkland	25
Dry steppe	25
Temperate steppe grassland	15
Main Taiga	25
Lakes and open water	5
Ice sheet and other	100

The package *gdistance* has originally been developed for the use of elevation model (Van Etten 2017). It was shown previously that costly areas, such as steep terrain, are assigned by a low value of speed. The function in *gdistance* is based on this assumption. Thus, the costs must be reversed; as in, for instance, easy crossing cells for *alpine toundra* are assigned by high values.

3. Transition Layer

The method of accumulative costs is based on raster data, in which each cell has been assigned by a specific cost value. In terms of route calculations, the raster data must be transformed into graphs where the centers of the cells are connected. The transition between two cells is shown through these graphs. The conversion to the graph data is called *transition layer*. To explain, the transition layer defines the actual costs of moving from one cell into an adjacent cell and is generated separately for each layer. The most important parameter is the possible direction of traveling. Figure 17 visualizes some different possibilities. On the left, the movement is only possible horizontally or vertically. This is defined by a neighborhood of four. The next option allows movement in the diagonal direction. Thus, from each point, all eight neighbors can be reached. The next level would be a neighborhood of 16 and so on. As in most of the GIS model, this analysis uses a neighborhood of 8, so that traveling can be done horizontally, vertically and diagonally. Increasing the neighborhood may enhance accuracy, but, on the other hand, it clearly decreases efficiency.

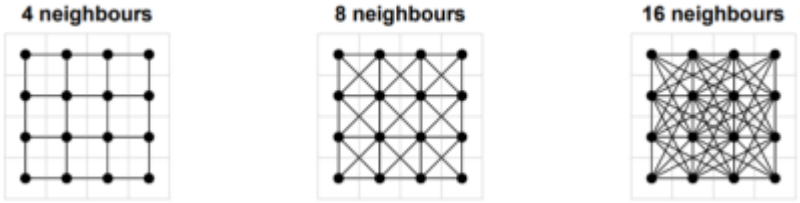


Figure 17: Definition of reachable neighbors. 4 Neighbors mean only horizontal and vertical movements, 8 neighbors allow also diagonal moving and 16 the movement to all 16 neighbors (Van Etten 2017)

The graphs of transition are assigned with weights that include the actual costs of the cells. The weights represent the probability of transition and are called permeability. To clarify, permeability is the mean of the inverse of the resistance. For instance, two cells within a steep area have got high resistance and thus a low permeability (Van Etten 2017). The exact permeability is defined by the inverse of the mean costs of two adjacent cells:

$$12)\ per = \frac{1}{mean(accCost_{AB})},$$

where *per* defines the permeability and the resistance, it can be defined as the costs of moving from one cell into another as shown in equation 12.

4. Merge Layers

After the assignment of the costs for each layer and calculating the transition costs, they must be combined in the next step. There are some technical and theoretical issues to consider. One technical aspect is that all layers must have the very same properties. First, all vector layers must be converted to raster objects. Second, all raster objects must contain the same resolution and bounding box. The theoretical aspect includes two important objectives: i) the transformation of the costs and ii) the weighting of the layers.

Cost Transformation

The necessity of the cost transformation can be explained through the example of the hiking function. The function describes the speed of human beings in different topology. Thus, the function describes a phenomenon on a high resolution. The resolution of the elevation model lies to 700 meters. Therefore, a smoothing effect of the human speed occurs as visualized in Figure 18.

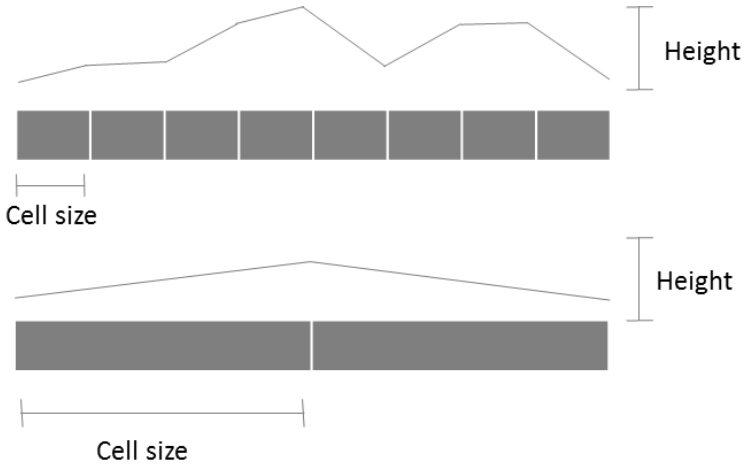


Figure 18: Smoothing effect of the slope due to resolution limitation

As a consequence, extreme values are lost and the effect of smoothing occurs. This can be partly corrected through the power transformation of the cost values from equation 9:

$$13)\ cost_{AB}(DHM) = cost_{AB}(DHM^{a}),$$

where *a* describes the variable of the cost transformation. Previous studies have neither dealt with nor discussed this issue. Thus, this work considers different possibilities of cost transformation and evaluates their impact on the definition of accessibility.

Weight

To compute accumulative cost surfaces, all transition costs must be combined into one. Here, there is the possibility to weight the layers differently for the assumption that some layers are more crucial for human movements. Correct weighting is only possible if all costs are within the same range; in other words, they must be normalized. When combining equation 9 to 13, the following formula emerges:

$$14)\ accCost_{AB}(DHM, Surf) = b_1 * cost_{AB}(DHM^{a}) + b_1 * cost_{AB}(Surf),$$

where moving from one cell A into cell B based on two layers $accCost_{AB}(DHM, Surf)$ is accumulated through the costs based on the elevation and the historical surface, where *a* describes the power transformation and b_1 describes the weighting of the elevation *and* b_2 of the historical surface.

To date, researchers have not treated the question of weighting in much detail and it is not clear if b_1 *and* b_2 must be different or the same. Therefore, this analysis opens with the assumption of equal weighting as suggested by Howey (2011). Same weightings are reached through an assignment of

0.5 for both variables. However, because to date, scientists have mainly used the elevation model, a higher weighting of this data should be considered. Some sensitivity analyses describe slope as the dominant contributor. The impact of doubling the transition cost of the slope has been five times higher than doubling any other factor (Todd & White 2009). Because of this insight and due to missing scientific observations, different weightings are going to be tested within this thesis.

5. Geocorrection

The accumulated costs are not only dependent on the cost surface, but also on the distance. To obtain correct distances, a *geocorrection* is necessary. This function takes two properties into account. First, it makes sure that diagonal traveling is costlier than vertically or horizontally traveling as shown in Figure 19. Second, the correction is due to the projected reference system. In WGS 84, the actual distance of one degree on the equator is smaller than one degree next to the poles (Gimond 2017).

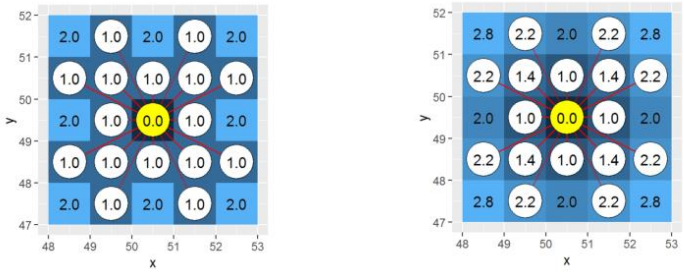


Figure 19: Cost surface on the left without Geocorrection. On the right, the actual distance has been included in the costs. (Gimond 2017)

6. Accumulative costs

So far, the transition layer includes all actual costs from moving from any cell into an adjacent cell. These values can be accumulated for specific routes. In this analysis, the costs of all routes from one origin in any direction are computed. An example is shown in Figure 20. The cost of the origin is zero and increases when moving in any direction and by the specified costs. The higher the accumulative costs, the more time is needed to reach the original point.

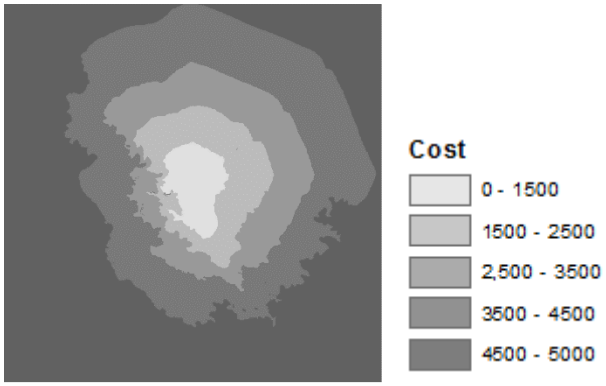


Figure 20: Accumulative costs from one point in all directions. The costs represent the degree of reachability.

7. Reachability Polygons

The accumulative costs can be described as the energy that goes into moving. Thus, with a specific amount of energy or specific amount of days to walk, people can reach a wide area in which traveling is easy like flat grassland. On the contrary, when using the very same amount of energy, the possible area that can be reached is much lower for regions that are hard to walk through, such as steep forests. Technically, this threshold t of energy can be defined as the maximum *max* of the accumulated costs $accCost_{AB}$ from equation 14:

$$15)\ t = max(accCost_{AB}(DHM, Surf)).$$

This threshold creates a *reachability polygon* as visualized in Figure 21, where the size of the reachable area within a threshold of costs does represent the average accessibility around each language point in this environment. In both examples, the region north- and southwards is better accessible than in the direction of west or east. However, the area of the reachability polygon on the left is larger than the one on the right. Consequently, the language point on the left is defined as less isolated as the language point on the right.

Figure 21: Cutting the accumulative cost surface based on a maximum threshold, in this example 2500, results in a reachability polygon.

5.2.3 Model Area of Contact

At this point of the analysis, a method to measure the degree of environmental isolation has been developed. This is a first estimation of the probability of contact with other communities. However, the neighborhood has not been considered yet with regards to using this approach. Within the method of IBD, the distance to the nearest neighbors has been computed. Now, the goal is to measure the possibility of contact on the recent multi-criteria surface. This is done by determining the overlapping areas of the reachability polygons. This overlapping part will be called the *area of contact* and is visualized in Figure 22 as the red area.

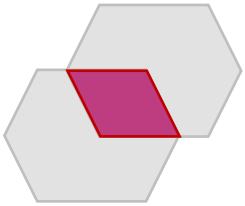
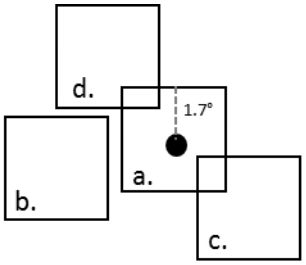


Figure 22: The overlapping part of two reachability polygons (red) is called the area of contact. (Mönkeberg)

Using a brute-force method, each reachability polygon could be compared to all other polygons. The computing costs would then be n*n with n the number of languages. However, an approach that uses spatial raster data would result in an immense cost of time. Thus, in a first step, we reduce the number of polygons that are compared. The next chapter describes the method to define the possible neighbors with a simple approach.

1. Overlapping Extents

There is a simple way to define the possible overlapping neighbors of each language points. It is based on the simple comparison of the extents. For each point, an extent of the same dimensions is defined as visualized in Figure 23. For example, for each language point, a bounding box (BB) of 1.7° in each direction of the original x and y is created. The size of this extent is dependent on the set threshold of accumulative costs while the polygon of reachability must be fully included in it. The determination of the overlapping parts of the extents is computationally cheap. The result is a list of all overlapping pairs of extents. In the example below, the BB of a is overlapping with the BB of c and d.



Overlap. 1	Overlap.2
a	c
a	d

Figure 23: Creating bounding boxes of the same extent for each language point. All overlapping extents are then summarized within a table as shown on the right.

2. Area of Contact

Now that all possible contact neighbors have been defined, the actual polygons of reachability can be compared as visualized in Figure 23. As a result of the previous step, only the polygons of possible neighborhoods need to be compared. With an average number of 4 neighbors, the computing costs are 4n, which is more than 50 times faster as the brute-force method. Similarly to the overlapping extents, the spatial polygons are checked for overlapping areas. All overlapping areas from point A to any neighboring language point are merged, which is also shown in Figure 23. This results in a total area of contact for each point.

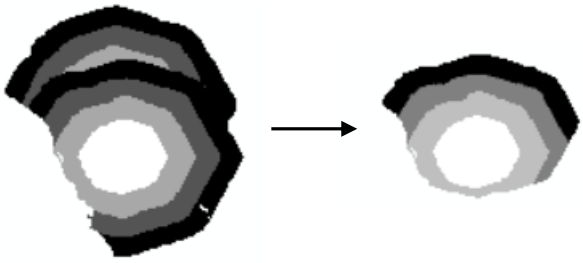


Figure 24: To the left, the actual reachability polygons are compared and the overlapping part (to the right) determined the area of contact.

5.3 Language and Isolation

Research question number 4 is about determining the impact of accessibility to language particularity. This will be done based on the created variables from chapter 5 and is visualized in Figure 25. Particularity has been defined as genealogical, global, focal and local PI. The degree of Isolation and reachability is reflected respectively within the variables of the reachability polygons and the area of contact. In order to address research question number 4, the relations between the variables of particularity and the variables of geographical reachability are determined. In chapter 6, it will be shown that the variable of local PI is not appropriate in terms of the correlation analysis.

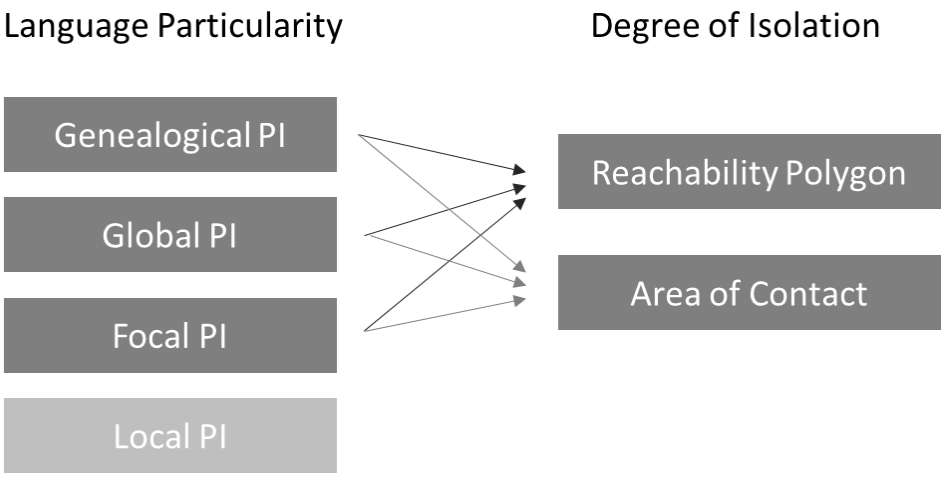


Figure 25: Visualization of the regression analyses that will be conducted based on the created variables to define language particularity and the degree of isolation.

The correlation to the reachability polygons focuses on the impact of the different parameters that have been explained in section 5.2.2. These are namely the impact of the cost transformation, the weighting of different layers and the definition of the threshold of the accumulated costs. Through regression analysis, the model that is able to explain language particularity the best can be evaluated. In all regression analyses, specific attention is paid with regards to the variation or spatial dependencies of the subcontinents.

6 Results

This chapter summarizes the important results of all analyses presented in *Chapter 5* Methodology. The results are structured as in the previous chapter. This means that, to begin with, the results of the genealogical and typological particularity are presented. Thereby the focus lies on the impact of the change in the different scaling-levels. After, the relation of the IBD and the altitude to language particularity are shown in section 6.3. In the section *Reachability,* first the verification of the model will be done. In addition, it will present a sensitivity analysis for the unknown parameters. In a third step, the correlation of the areas of reachability to language particularity is shown. The last section *Area of Contact* shows the results of the extended model of reachability. Thereby, the correlation between the area of contact and language particularity is analyzed. The meaning and interpretation of the results are treated in chapter 7.

Due to the weak correlations between most of the variables, this chapter focuses on the qualitative description of the relations instead of statistical regressions or significance levels.

6.1 Genealogical Particularity

The result values of the genealogical PI are visualized in Figure 26. The values range from 0.5 to 1 with a mean of 0.84. Only one-quarter of all values are smaller than 0.76. To put it simply, most of the families contain few languages and are thus adjusted by a high particularity value. Only a minority of the languages owns many languages and is assigned to a low particularity value.

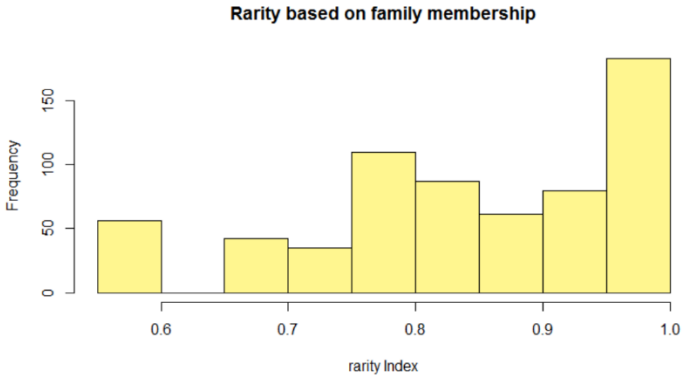


Figure 26: Value distribution of the genealogical particularity

The distribution of the rarity values gets more interesting by adding the spatial component as visualized in Figure 27. The spatial distribution presents some clear patterns. While the first quarter of the values does not seem to be very meaningful in pure statistical terms, their spatial distribution shows clear patterns. There is not a single language in South America with a rarity index smaller

than 0.7. Interestingly, South America seems overall to be more particular with regards to family membership than North or Central America. Furthermore, South America seems to be slightly more particular on the west than on the east. Central America, on the other hand, represents the opposite. The two lower categories clearly cluster in this region and in Mexico respectively. In addition, there are other clear patterns in Central America. Around Guatemala and to its south, a medium zone of particularity, a noticeable zone of higher particularity can be defined. North America represents the highest diversity of language particularity. In the southeast as well as in Alaska, a region of high particularity can be observed. There are mixed values on the western coast of northern America, nevertheless, it does not seem to be a random pattern.

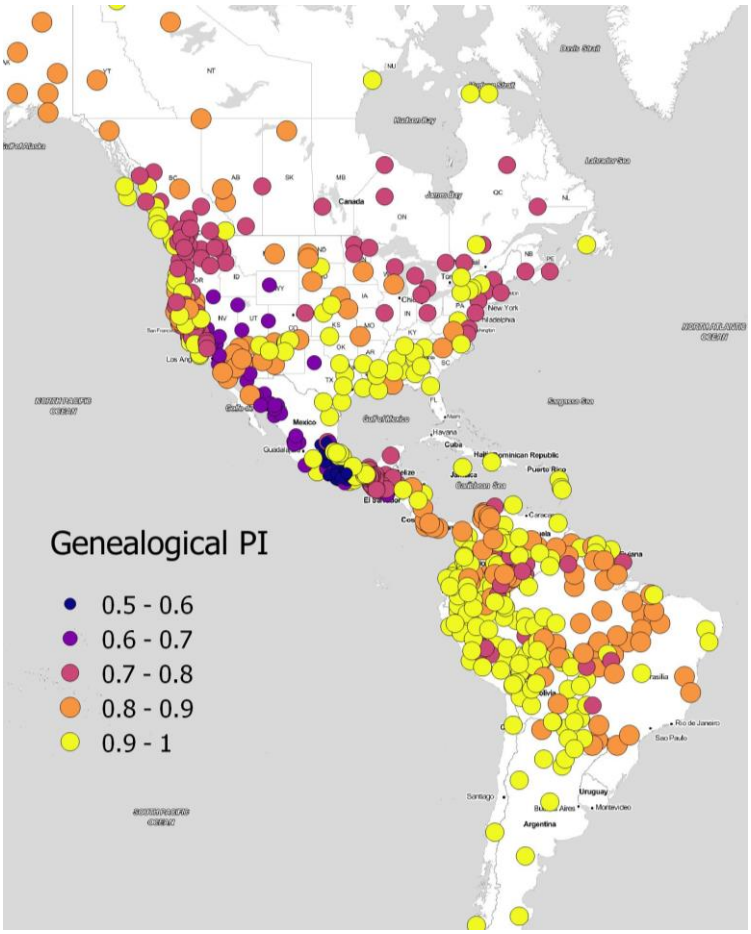


Figure 27: Spatial distribution of the genealogical PI, categorized and visualized in different colors. Parts of Central America are shown in higher zoom level to the left.

It is important to understand the meaning of the particularity values. Languages of the same family always share the same value of particularity. However, languages sharing the same value of particularity do not automatically belong to the same family.

6.2 Typological Particularity

As explained in chapter 5.1, this analysis has first considered the weighted mean particularity index, which is suggested by different scientists (Daumé 2008, Georgi et al. 2015). As shown in Figure 28, the typological particularity is mainly influenced by the count of missing information instead of by the information about the features. The reason for this lies in the sparse information of the language dataset. Thus, this index does not reflect typological particularity, but the number of available information, and, consequently, is not appropriate for these analyses. As a result, the particularity index suggested by Cysouw (2017) has been determined and used for all further analysis.

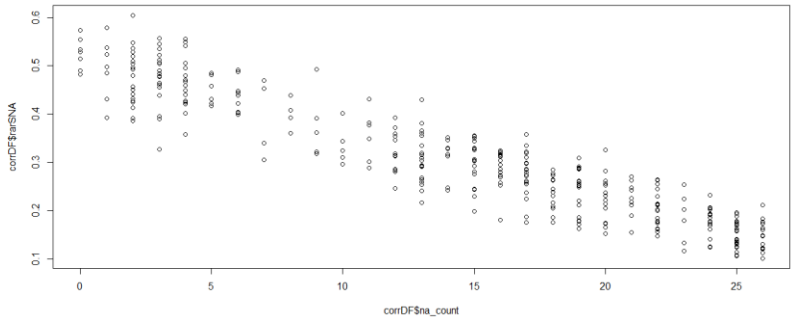


Figure 28: Relation between the number of na's of a language on the x-axes and the focal PI on the y-axes

The typological particularity by Cysouw (2017) has been computed for three different geographical sample sets. As explained in chapter 5.1.3, the global PI includes the whole study area, the focal PI has split the dataset into North and South America and the local PI defines particularity based on the neighbors within 1000 km. After the data reduction suggested by Daumé (2008), as explained in the methodology, there are 342 indigenous languages included in the study area of the Americas. Importantly, 210 languages exist within South America and 112 within North America.

The high impact of language particularity due to scaling can already be shown in simple statistical visualizations as Figure 30. Overall, the range of the values is much higher than for the genealogical PI. While the values for the global PI seem to present a normal distribution, the count of very particular languages increases for local PI. It seems that the smaller the subset, the more likely languages are identified as particular. This tendency draws attention. It was already said that the local PI might risk the enhancing of number of na's. Thus, the comparison of two languages lies within very few features. The value distribution supports the suspicion that the local PI cannot reflect the actual particularity, but is biased through missing information. Thus, this index will not be considered to determine correlations between language particularity and geographical isolation.

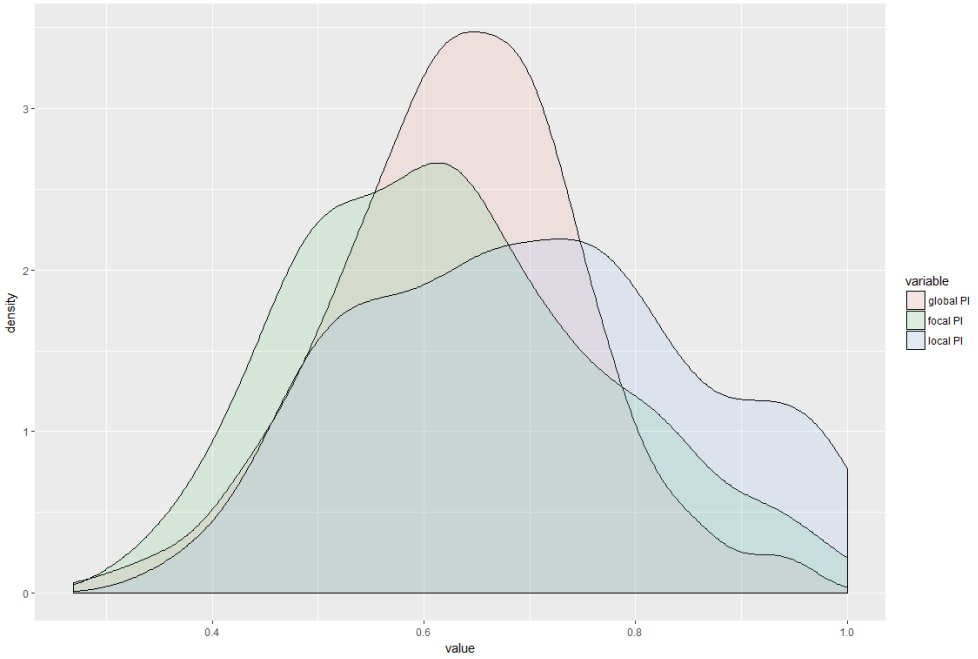


Figure 30: Distribution of the TP values of different scaling: left: global, middle: focal, right: local

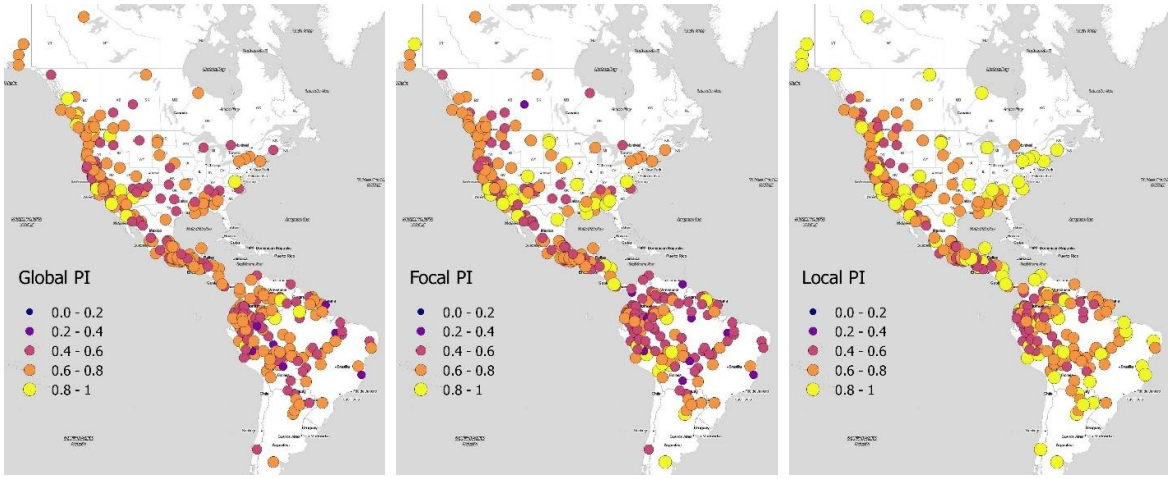


Figure 29: Spatial distribution of the typological particularity indices: left: global, middle: focal, right: local

Figure 29 shows the spatial distribution of the indices. Overall, the distribution of the global PI seems continuous. All categories can be found more or less in the whole study area. The exception may be the category of the smallest PI, the majority of which belongs to South America. A very similar scattering can be found for the values calculated on the base of the focal level. However, there are some differences. For example, the particularity in North America increases for many languages in the focal PI. The middle image shows two zones of particularity. The middle east coast can also be observed for the global PI. The second particularity zone on the western coast is unique for the focal PI. In South America, the focal PI seems to be decreasing in general. Especially the languages along

the eastern cost tend to be classified as more common. In addition, there are more languages of the most common category for the focal PI in South America. The local PI on the right image does not show any clear patterns. However, an increasing of particularity over the whole study area can be observed.

The exact differences between the three scales are visualized in Figure 31. On the left image, the difference between the global PI and the focal PI is shown. There is a clear pattern of a decreasing particularity within South America and an increasing of the values for North America. In other words, languages in South America tend to be more common in case we only compare them to the South Americas. Moreover, languages in the north tend to be more particular when only compared to the North American languages. The interpretation and opportunities of this insight will be discussed in 0. The middle image of Figure 31 shows the difference between the local and the focal PI. The most remarkable change is the high increase of particularity of a few languages visualized in yellow. In contrast to the left image, the languages in South America are increased in average in their particularity. Small or no change can be observed along the western coast for both subcontinents. The last image shows the change of the global to the local PI. There are no clear patterns or clusters of the change.

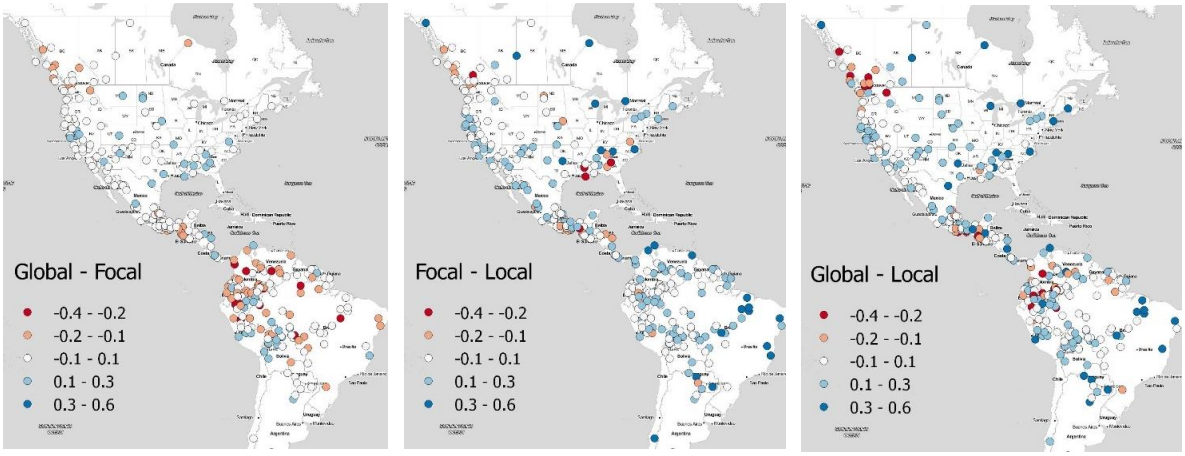


Figure 31: The change of the TP values, due to scaling: left: global to Focal, middle: Focal to Local, right: Global to Local

Relation of genealogical to typological particularity

Language particularity can be described based on two methods - genealogical or typological particularity. Many scientists claim that languages within the same family are more similar compared to languages of other families. In light of the methods of this analysis, the assumption is not true. Figure 32 plots the relation between the genealogical and the global PI. There is no clear

correlation that would indicate that family belonging reflects typological properties. This means that the choice of the measurement is crucial to gain insights about language particularity.

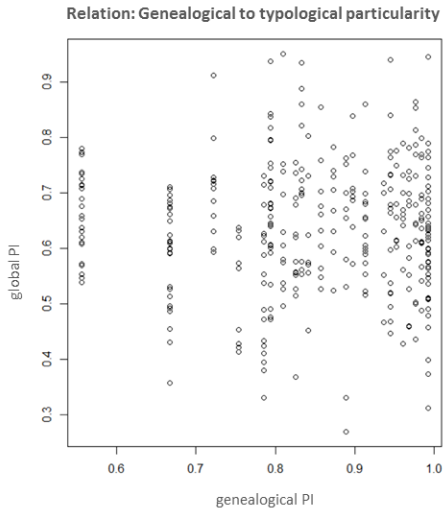


Figure 32: Relation between genealogical PI and global PI

The local PI was able to underline the relativeness of the indices. However, as already pointed out, the local PI is likely to be influenced through missing information and will not be considered for correlation analyses due to geographical reachability.

6.3 Simple factors of Isolation

The state of art has shown that simple factors of isolation can explain language particularity. However, the applied methods are mostly non-reproducible. Additionally, none of the analyses has been conducted for the same study area and data set. As explained in RQ 2, a first analysis will show if and how these insights can be transferred onto the framework of this thesis.

IBD

Isolation can be understood in terms of metric distances as suggested by the concept of 'isolated-by-distance' (IBD). One metric measurement for isolation is, for instance, the number of neighbors within a specific radius. The assumptions of this thesis would expect a high language particularity for all languages within a small neighborhood and a small particularity for languages surrounded by many neighbors (see Figure 33).

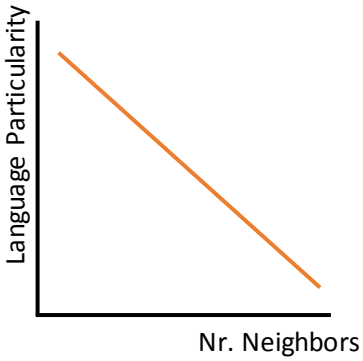


Figure 33: Expected correlation of language particularity and the number of neighbors within radius r

The main question for this simple method is the definition of the radius r. Research on this topic has not touched upon this specific issue. Thus, this analysis has compared the radius of 50, 100, 300, 500, 1000 and 2000 meters to define the threshold of isolation. For each radius, the correlation between the number of neighbors and the three different particularity indices has been defined. Not all correlations correspond to the expected outcome. The typological PI did not show clear relations to either of the radii. The genealogical PI has shown dependency to all radii. The strongest relation can be found for a radius of 100 and 300 meters. Using a radius of 300 meters, the frequency of the number of neighbors is continuously decreasing as visualized in Figure 34.

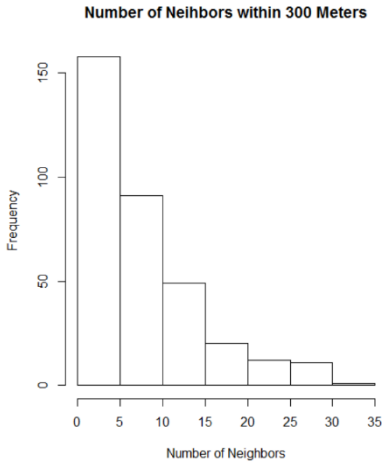


Figure 34: Histogram of the number of neighbors within a radius of 300 meters

Figure 35 shows the relation of the number of neighbors within 300 meters to the global PI in the left image, the focal PI in the center and the genealogical particularity to the right. The two typological PI do not show any clear relation to the neighborhood. Although, some languages seem

to verify IBD using the genealogical PI. The language group in the right corner, for example, indicates that a large neighborhood does decrease particularity. Similarly, consider the dense point group in the upper left corner where high particularity is correlated to a high degree of isolation.

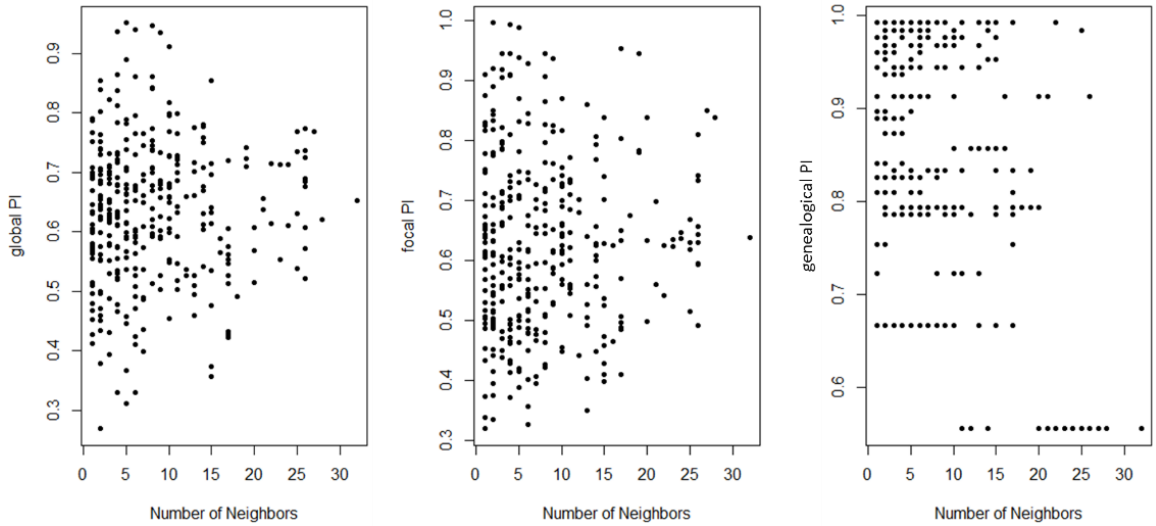


Figure 35: Relation of particularity to the number of neighbors within 300 meters. Left: global PI, middle: focal PI, right: genealogical PI

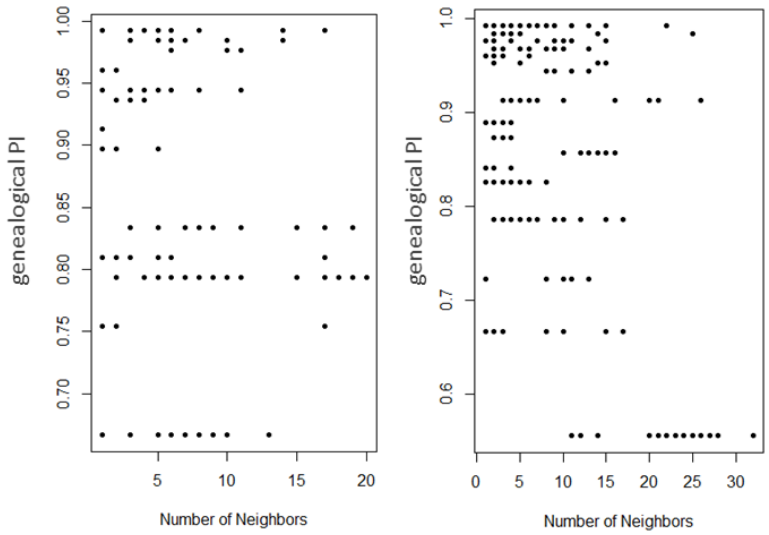


Figure 36: Relation of genealogical PI to the number of neighbors within 300 meters. Left: North America, right: South America

The next Figure 36 will show that the verified IBD above for the Americas does in fact reflect the IBD for South America on the right. Note that the range of particularity as well as the number of neighbors is smaller for North America. Thus, the whole language group of large neighborhood and small particularity is located in South America. While some of the expected correlation can be observed in the southern subcontintent, North America does not seem to fulfill IBD.

Altitude

The second assumption, that needs to be checked, is whether altitude is an indicator for language particularity. As altitude is assumed to be positively correlating to isolation, an ascending gradient between altitude and language particularity is expected (see Figure 37).

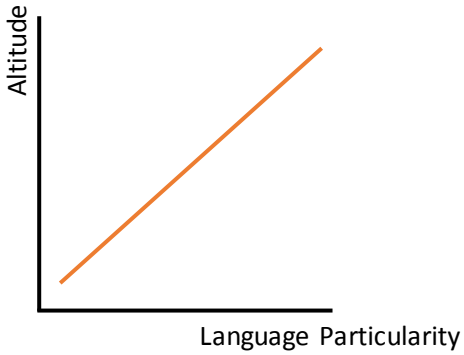


Figure 37: Expected correlation of language particularity and altitude

Based on the very different results in the IBD-analysis, North and South America will be treated separately for further analyses. The results in terms of the correlation to the altitude are very similar to North and South America. In fact, there is no significant correlation to neither of the typological indices. The genealogical PI shows the strongest correlation to the altitude (Figure 38). However, the correlation is not as expected. Instead of a positive correlation, high altitude tends to a lower language particularity. This can be observed in the point group of high particularity in South America, which shows small values of altitude and a slight density of low particularity for high values in altitude. Both points assigned by the lowest altitude range from low to high particularity.

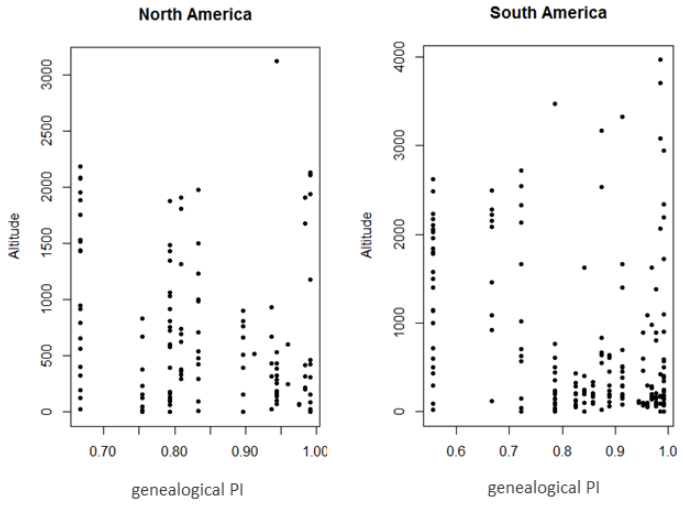


Figure 38: Relation between altitude and genealogical particularity, left: North America, right: South America

6.4 Reachability

As pointed out earlier, this thesis models isolation, respectively accessibility based on accumulative costs. Before analyzing the results, it is important to ensure that the implementation leads to the desired outcome. Thus, this chapter will first do a validation of the model. Furthermore, it will determine the impact of the uncertain parameters such as the transformation of the costs, the weighting of the layers and size of the threshold in form of a sensitivity analysis.

6.4.1 Validation

The examples below show that the model produces accumulative costs and that the outcome is reasonable for each input data. This means that the accumulative costs are higher for costly and heavy cells to move through as for cells that have been categorized as easier to travel through.

Elevation

Considering human movement only within topography, steep areas should result in higher costs than flat areas. Figure 39 shows that the language *Kawaiisu* is located within a very rough area in North America. The hillside to the left is much steeper than the terrain in the direction of the south. The terrain is reflected within the accumulative cost surface. The expected result of a less difficult traveling in the direction of north-south compared to traveling within rougher area to the west is fulfilled. It is noteworthy to mention that cells containing the sea are not included in the accumulative costs. This behavior of the model was observed for the whole study area and the method to compute accumulative elevation cost surfaces could be verified.

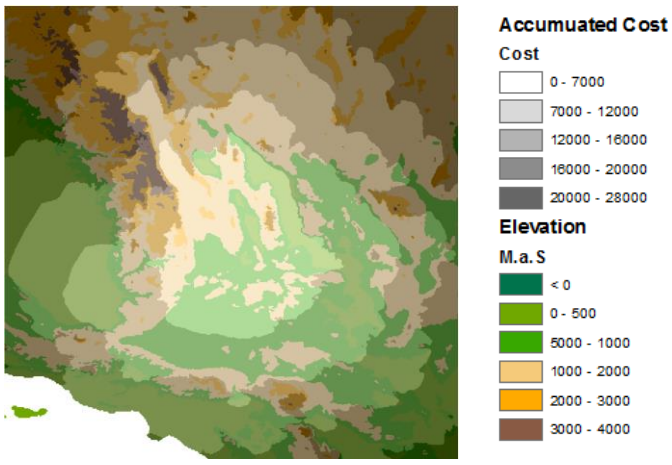


Figure 39:Accumulated costs based on the elevation around the point Kawaiisu

Historical land surface

The same logical verification of the method must be done for the historical land surface. Surfaces that are easy to travel through must have lower accumulative costs than expensive cells. The language *Cholon* is situated on a mountainous area near the coast. The historical surface is mapped in Figure 40. The most costly surface is represented by the *ice sheet and other permanent ice*. The *montane mosaic* is an easy surface to travel through. The surface along the coast is less costly and the surface in direction of the east is the cheapest in this area.

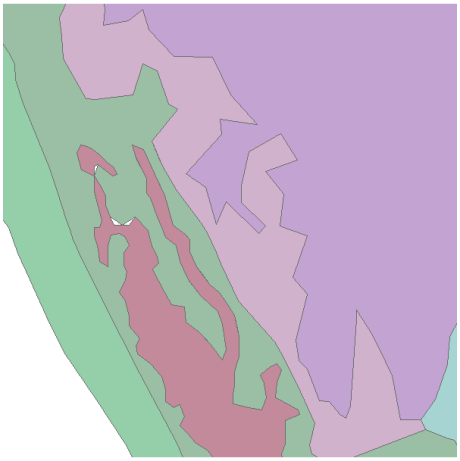


Figure 40: Historical Surface around the language point Cholon. Blue: Ice sheet, Turquois: alpine mosaic, yellow: tropical extreme desert, green: moderate tropical forest, orange: tropical rainforest

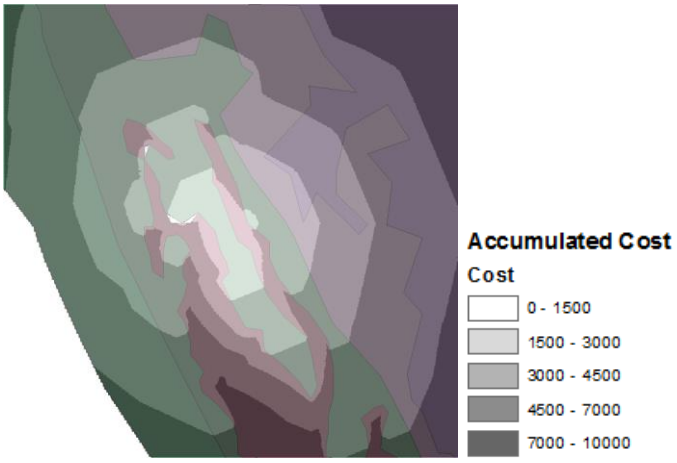


Figure 41: Accumulative costs based on the historical surface around the language Cholon

In Figure 41, the surfaces are transformed into costs. *Montane tropical forest* and *tropical rainforest*, for example, share the same and lowest value assignment, followed by the coastal surface. Note that the dark green one and the highest costs have been assigned to the brown one. The accumulative cost surface reflects this pattern. The pattern of the most costly surface is reflected as a warped cost surface.

Also, it becomes clear that traveling to the east is easier than traveling to the west. As a result, the accumulative cost surface is correct.

6.4.2 Sensitivity Analysis

In chapter 5.2.2, the three parameters of power transformation have been defined as variable *a*, the weighting as variable b_1 and b_2 and the threshold of costs have been described, too. As already explained, there is a clear lack of scientific insights for the exact definition of these parameters. Therefore, it is important to know how sensitive the model responds to them. The sensitivity analysis determines the impact of the modification of the parameters with regards to the output of reachability. This will determine whether the change of some parameters have significant influence to the result or not.

Cost Transformation

In previous chapters, it has been explained that due to resolution effects the Tobler's hiking function may not fully accord with this analysis. In terms of the smoothing effect, it may be reasonable to empower the value of the speed based on elevation. The detailed explanation can be looked up in section 5.2.2. The impact of transforming the value of the elevation value will be shown in the example of the language *Campe (Axininca)* in South America. Figure 42 maps the terrain around the language point. The language point itself lies within a flat area and at the beginning of a valley. To the east and south, there is a very mountainous and steep area.

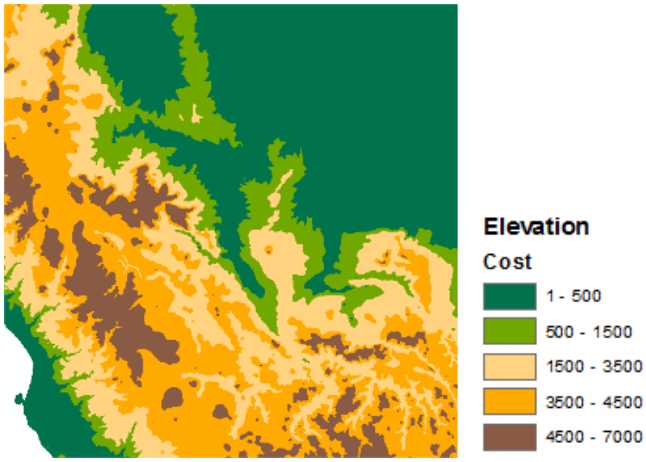


Figure 42: Elevation model around Campa (Axininca)

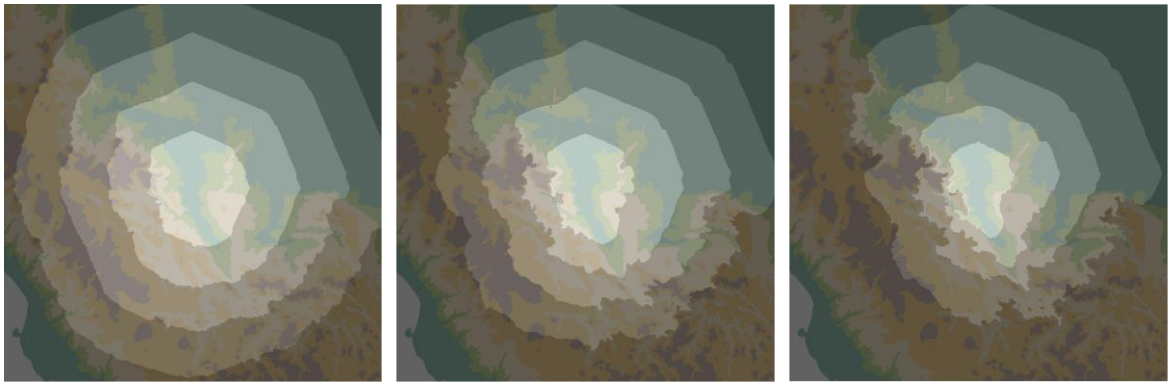


Figure 43: Accumulative cost surfaces based on different power transformation of the elevation model. Left: a=1, middle: a=2, right: a=3

Figure 43 shows the different accumulative cost surfaces for various power transformations of the hiking costs. The first one represents the value of Tobler's hiking function. For the second one, these costs have been empowered by 2 and in the third picture by 3. Through the normalization of all values between zero and 50, they can easily be compared to each other. As expected, the steep area in the southeast becomes more and more costly. On the contrary, the valley to the south becomes easier to travel in comparison to other terrains. The flat area in the north barely shows any differences.

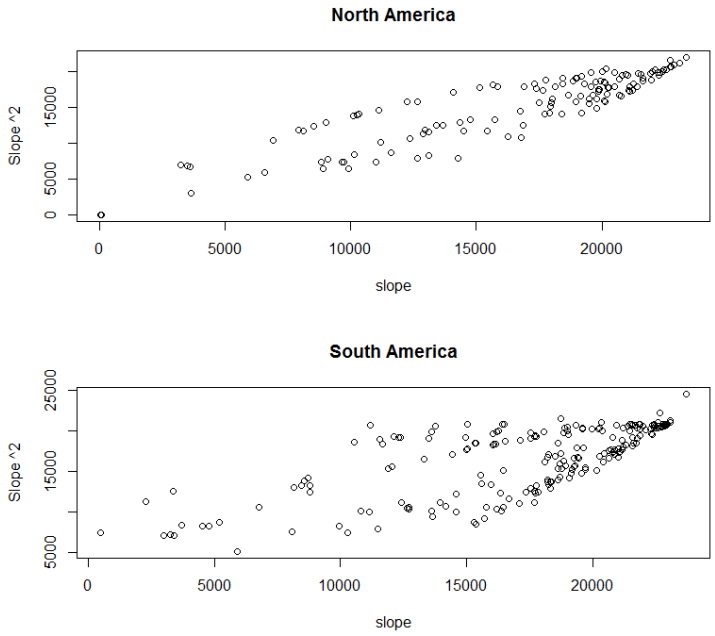


Figure 44: Relation of the reachability polygons by an empowering of the slope by 1 to the empowering by 2.

Figure 44 shows the relation between the reachability polygons produced through different power transformations of the elevation costs. Overall, the change of the scaling shows high linearity. Additionally, languages in flat areas and thus with high reachability are less sensitive to

the power transformation of the slope. However, the impact of languages with low areas of reachability is less regular, especially in South America.

Weighting

Another important but poorly determined component is the weighting of different layers. The impact on the change of weighting will be shown in the example of *Ignaciano* in South America. Figure 45 shows the two input layers to calculate reachability. The terrain around the language point is flat and easy to travel through. In the southeast, there is the onset of a rather mountainous area. The historical land surface splits the flat area into two categories. Considering the costs as speed, the point of reference for the language in question lies within a very easy land surface (yellow) to move onto, but at the same time it is located near a border (orange) that slows the speed down. The mountainous area is recognizable as an even slower speed. Thus, the weighting of the two layers is crucial to define the accessibility around this language point.

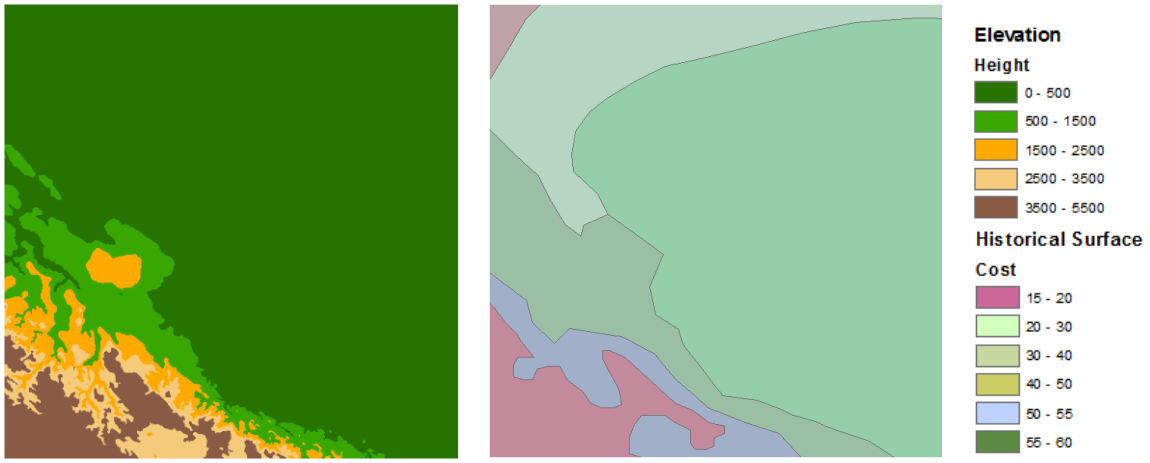


Figure 45: The terrain (left) and historical land surface (right) around the language point of Ignaciano.

The sequences of Figure 46 show the impact of changing the weight of the layers slope and historical surface. Moreover, the sequences show the change of the proportion of the two layers. The first one shows the proportion elevation to land surface with a ration of 1:0, the second one shows the proportion 2:1, the third 1:1 and the last one 1:2. It becomes clear that the higher the proportion of historical surface, the higher the reachability within the less costly surface in eastern direction is.

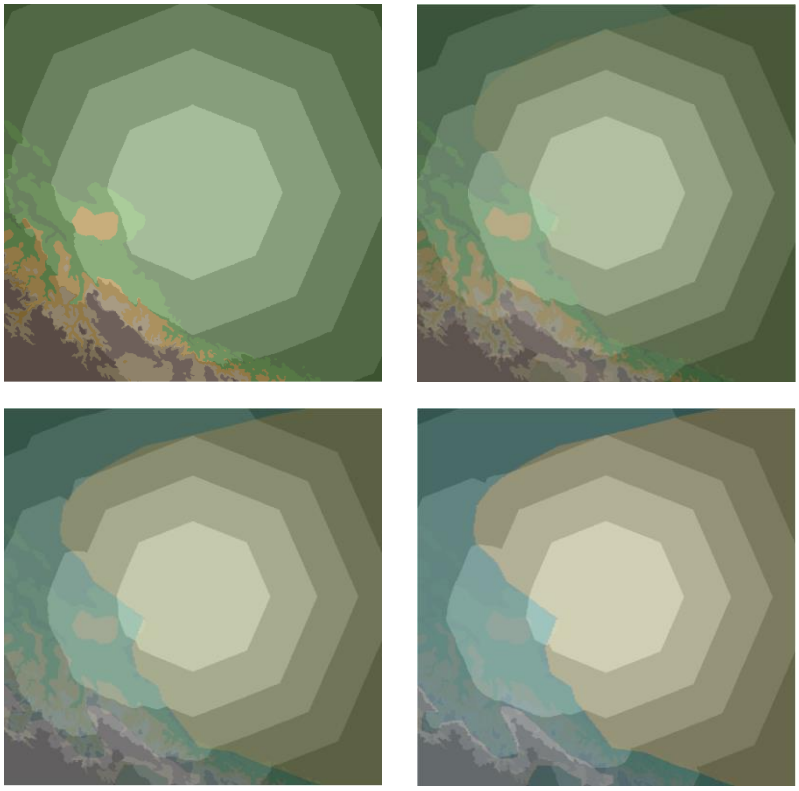


Figure 46: Accumulative cost surfaces by different weightings. Weights as [b_1 / b_2]: upper left: 1/0, upper right:0.7/0.3, lower left: 0.5/0.5, lower right: 0.3/0.7

Figure 47 visualizes the impact on the whole study area. Whereas the accumulative costs show a continous distribution of the reachability for a full weighting of the elevation model, the distribution is aprubt when taking into account the historical surface.

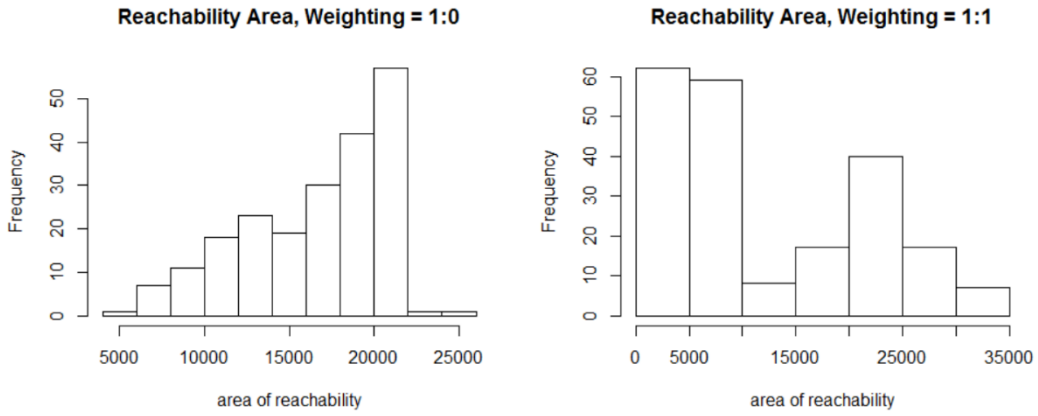


Figure 47: Value distribution of the area of reachability, by a threshold of 2500 and different proportions slope: veg: left: 1:0, right: 1:1

Threshold

There are no scientific insights about an appropriate threshold to measure the reachability of language points. Thus, it is important to determine the impact and sensitivity of this parameter. In order to answer how crucial the set of the threshold is in terms of the output, the model has been tested for different thresholds. The threshold defines how far humans can travel in each direction given a specific energy. Within an area of the same costs, as for instance flat croplands, an eight-angular polygon results. The distance to one edge then defines the maximum distance that can be reached. This polygon can be simplified with a circle and the radius represents the maximum distance.

To reach an average radius or distance of 72 km, which corresponds to a three day's walk, the threshold must be set to 2500. The sensitivity of the threshold varies depending on the weighting of elevation and historical surface. How does a doubling of the threshold influence the area of contact? The individual example in Figure 48 expresses the impact on changing the threshold of the accumulative costs. The area of reachability on the left is much smaller than for the threshold on the right, that is, twice as big.

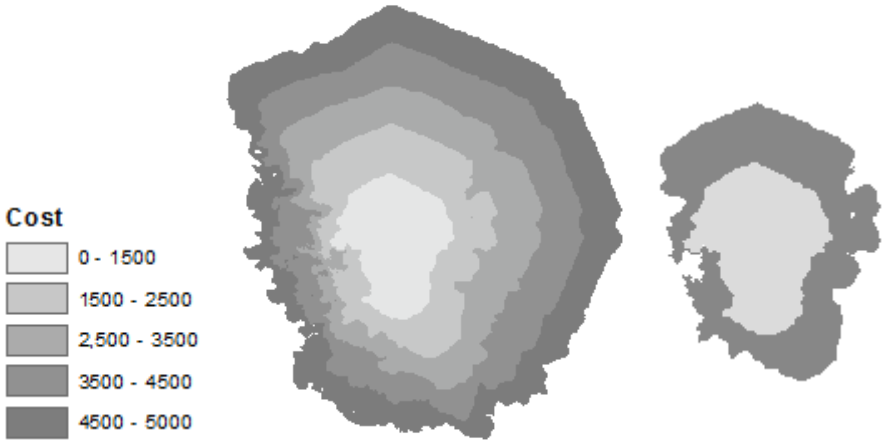


Figure 48: Impact of different settings of the thresholds of the accumulative costs to define the area of reachability. Left: 2500, right: 5000

Table 5 lists the statistical values to determine the impact of changing the threshold of accumulative costs. It is important to highlight that South and North America have been treated separately and models of different weightings have been examined. The *Range* presents minimal and maximal estimated distance in one direction that can be reached by the defined threshold. For example, in South America, the average radius of the reachability polygons lies between 40 and 87 km by a threshold of 2500 if we consider the elevation model only. The doubling of the threshold results in a range that is twice as big. Less clear is the impact of changing the threshold for North America and

a weighting of 0.5 for elevation and 0.5 for the historical surface. It seems that the bigger the threshold, the less linear the change of the reachability polygons is.

Table 5: Statistical values to determine the impact of changing the threshold for different models

Area	b_1:b_2	t	Range [km]
SA	1:0	2500	40 - 87
	1:0	5000	80 - 170
NA	1:0	1500	3 - 51
	1:0	2500	3 - 86
	1:0	5000	4 - 105

b_1:b_2	t	Range [km]
1:1	2500	18 - 100
1:1	5000	38 - 200
1:1	1500	3 - 64
1:1	2500	3 - 171
1:1	5000	3 - 213

It seems as if the models that have a higher weighting of the historical land surface are more sensitive. This has been visualized in Figure 49. On the left, the relation of the area of reachability for the model that only considers the elevation of the different thresholds is shown. On the right, the same areas of reachability are compared, although for the model that uses the same weight for both layers. The change of the areas on the right is less linear and there are much more outliers. The reason for this may be that the land surface shows more abrupt changes compared to the elevation.

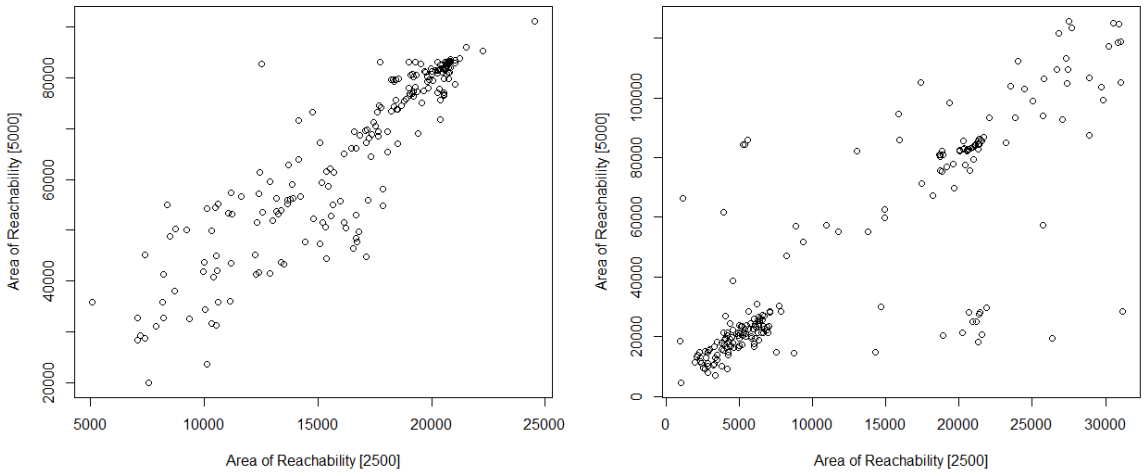


Figure 49: Sensitivity of the Threshold as the relation of the area of reachability based on the threshold of 2500 to the area of reachability based on a threshold of 5000. Left: b1 = 1, b2 = 0, right: b1 = 0.5, b2 = 0.5

Overall, the sensitivity analysis has shown different impacts on the uncertain parameters of cost transformation, weighting and threshold. The impact of the cost transformation on the elevation can visually be defined, but, generally speaking, does not highly influence the area of reachability. Thereby, North America is less sensitive than South America. Importantly, the definition of the weighting seems to be most crucial. Including the historical land surface significantly changes the result of reachability. The impact of the threshold has been determined as not too strong. In general, threshold and area of reachability are reacting mostly in a linear way. The impact is negligible especially with regards to continuous data. Moreover, outliers and non-linear relation are more likely for abrupt input layers.

6.4.3 Correlation

According to the hypothesis in RQ 3, the correlation between language particularity and the degree of isolation is expected to be positive as visualized in Figure 50. In a first approach, the degree of isolation will be measured as the area of reachability. The larger the area of reachability, the smaller the degree of isolation is; thus, we are presented with a negative correlation.

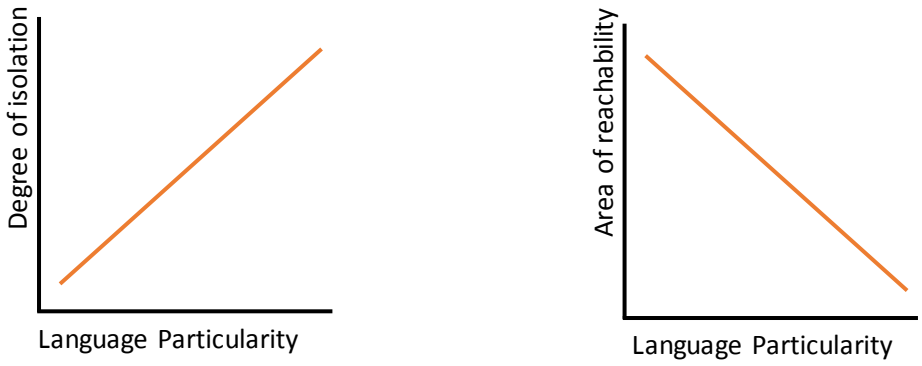


Figure 50: Expected correlation between language particularity to the degree of isolation (left), resp. the area of reachability (right)

The area of reachability will be computed based on the accumulative model. The sensitivity analysis has shown that the parameter of cost transformation, weighting and threshold can have crucial impacts on the model. Therefore, the area of reachability has been computed for different combinations of the parameter values listed in Table 6. The different outcomes have been tested for correlations to the genealogical, global and focal PI. In terms of computational limitations, the accumulative cost analysis has been conducted separately for North and South America. The results will show that this separation is reasonable or even necessary, because the insights on the two subsets are clearly differencing. The following section presents the evaluation of the best suitable

model for the different particularity indices. First, the results of the genealogical particularity and second, the results of the typological particularity are presented.

Table 6: Unknown parameters of the accumulative models and the values that have been evaluated

Parameter	Values			
a	1	2	3	4
b1	1	0.33	0.5	0.66
b2	1	0.33	0.5	0.66
t	1500	2500	5000	-

Genealogical Particularity

The best suitable parameters to explain genealogical language particularity for North America are the untransformed value of the elevation – that is, we only consider the layer of the elevation – and a threshold of a three-day's walk as summarized in Table 7. Similar to the northern subcontinent, genealogical particularity can be best explained through the weighting of 1 for the elevation in South America. However, the difference to the north is the cost transformation and the threshold.

Figure 51 shows the relation between the genealogical PI to the area of reachability. Thereby, the area of reachability has been computed for the models that are listed in Table 7. In North America, on the left image, the point group of low particularity is apparent. The point group represents the assumption that a high area of reachability explains low values of particularity. However, there are similar point groups of high areas of reachability that show higher particularity. Some more patterns can be observed in South America on the right. Surprisingly, there is a clear point cloud in the upper right corner. These points are controversial to the expected outcome and indicate that particular languages are located in well accessible areas. A similar trend shows languages of low particularity, which have rather small areas of reachability.

Table 7: Best suitable model for GP in North America

Area	*a*	*b1*	*b2*	*t*
NA	1	1	0	2500
SA	2	1	0	5000

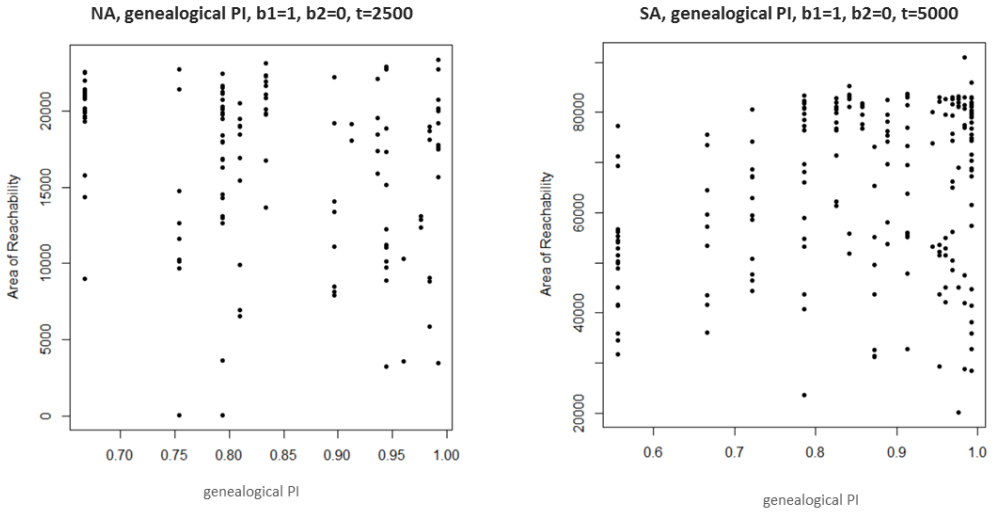


Figure 51: genealogical particularity and the area of reachability computed as listed in Table 7. Left: North America, right: South America

Typological Particularity

For North America, the best model has been evaluated as a simple scaling of the elevation, a weight of the historical surface twice as big as for elevation and a threshold of a six-day's walk. The clearest pattern was found for the global PI. In South America, the clearest relation has been determined for the focal PI. The model includes a cost transformation, a stronger weighting of the elevation and the threshold of a six-day walk. Both models have been summarized in Table 8. The relation of the area of reachability is visualized in Figure 52. No correlation could be found in either of the subcontinents. In North America, two clusters of language points can be observed. The point group at the bottom seems to show low reachability independently of the typological properties. All points assigned by the highest global PI are located within areas of low reachability and coincide with the expected outcome. The second point group is less clear and is represented by higher areas of reachability, but by lower values of particularity. Only few language points have a medium area of reachability. The distribution of the points in South America seems randomly arranged. The observed clusters in the left image can only be observed in weak form. Still, points of low values of the focal PI tend to have larger areas of reachability than point of higher values in focal PI. However, clear correlation cannot be observed in neither of the images.

Table 8: Best suitable model for TP in North America

Area	*a*	*b1*	*b2*	t	PI
NA	1	0.33	0.66	5000	Global
SA	2	0.66	0.33	5000	focal

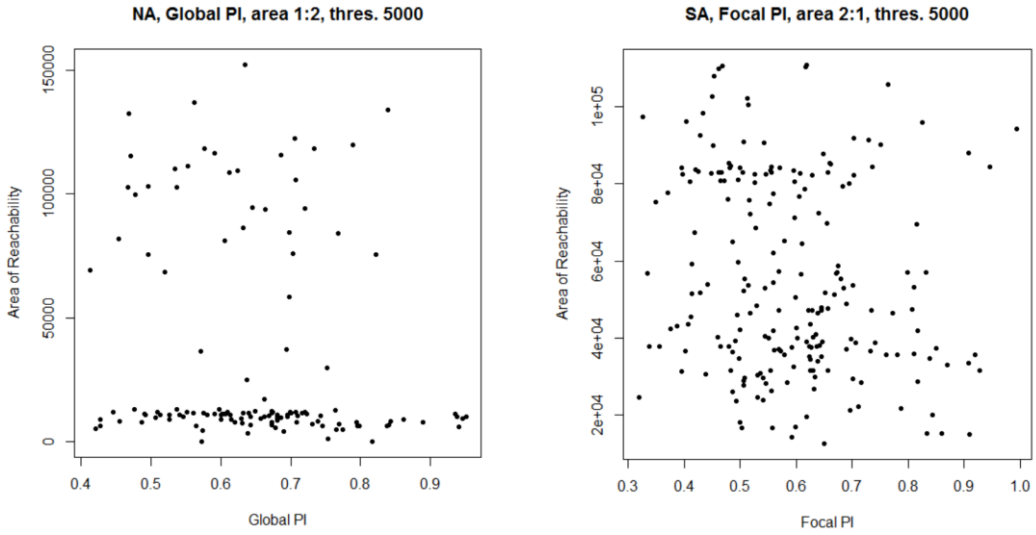


Figure 52: Typological particularity and the area of reachability computed as listed in Table 8. Left: North America, right: South America

6.5 Area of Contact

6.5.1 Validation

Figure 53 visualizes the area of contact in North America. The applied model does consider a power transformation of one, the full weighting of the elevation and a threshold of a six day walk. As expected, languages in dense regions are more likely to have a large area of contact. Nevertheless, the existence of small areas of contact shows the influence of the elevation model. Based on this logical verification, the model could be validated.

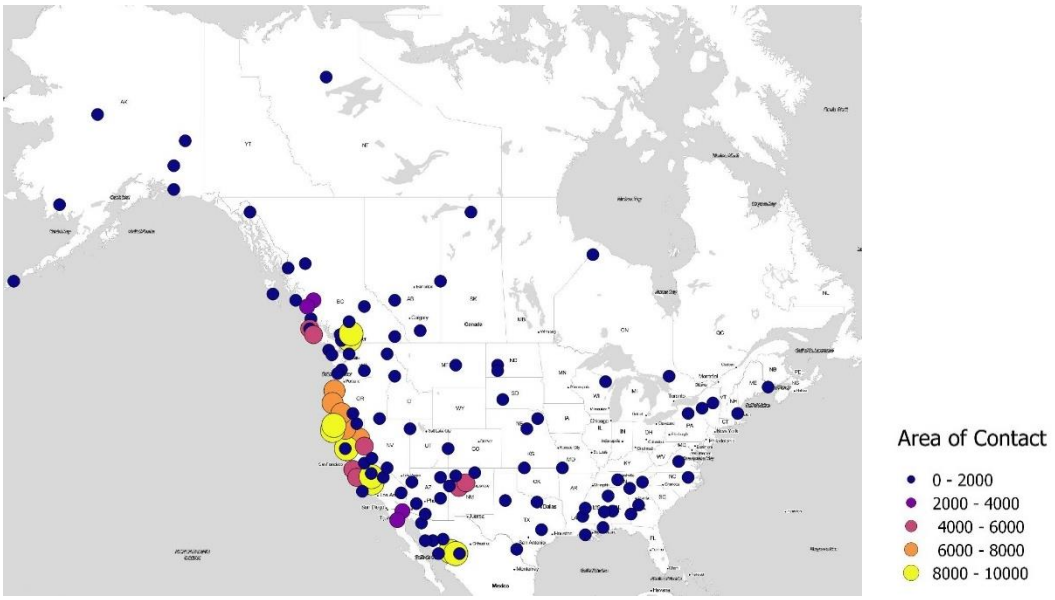


Figure 53: Visualization of the area of Contact in North America.

6.5.2 Correlation

The area of contact describes the overlapping parts of the reachability polygons. Similar to the correlation analysis of reachability, the area of contact has been determined by some different parameter. The combination of the parameters (cost transformation, weighting and threshold) leads to different results of the model. Due to the time-consuming analysis of the area of contact, not every possible combination could be evaluated. Thus, the areas of contact have only been determined for the models with the highest correlation between language particularity and areas of reachability as shown in chapter 6.4.3.

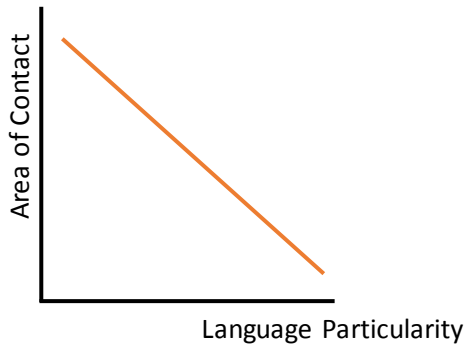


Figure 54: Expected Correlation between language particularity and the area of contact

The expected outcome of this analysis (Figure 54) is a clear negative correlation between the area of contact and language particularity. This section will evaluate the area of contact and the correlation to language particularity.

Genealogical particularity

Figure 55 plots the relation of genealogical PI to the area of Contact. Especially in North America, the few points that show an area of contact are not enough to find conclusions. In South America, the language points that show low genealogical PI are remarkable. It seems that large language families are independent of the area of contact. Nevertheless, there is an accumulation of points in the lower left corner. For them, the expected outcome of high particularity and low area of contact can be verified. Considering the results from the reachability analysis, the same model did show the opposite correlation.

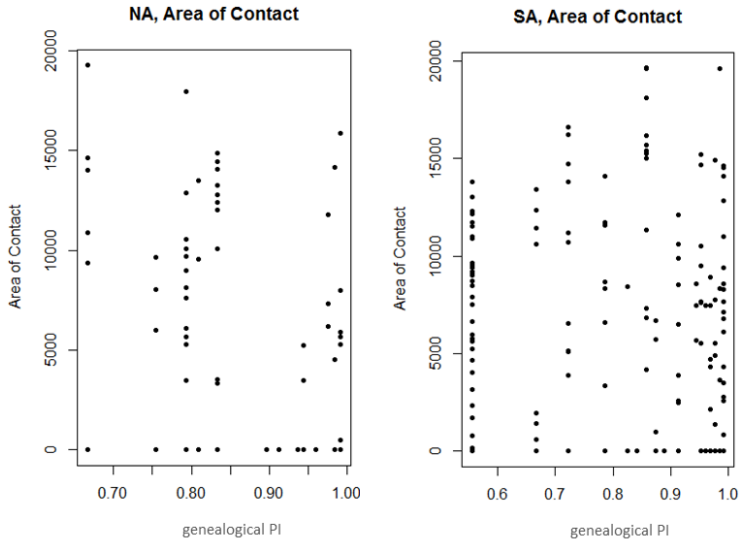


Figure 55: Relation between the genealogical PI and the area of contact. Left: North America, right: South America

Typological Particularity

For none of the models, clear relations between typological particularity and the area of contact could be determined as shown in Figure 57. In the plots of the genealogical particularity, the high number of zero area of contact has been concealed due to the abrupt values of the genealogical PI. However, the high amount of points with zero area of contact is obvious within the plots for the global PI. Unfortunatly, the rest of the language points does not leave much room for interpretation.

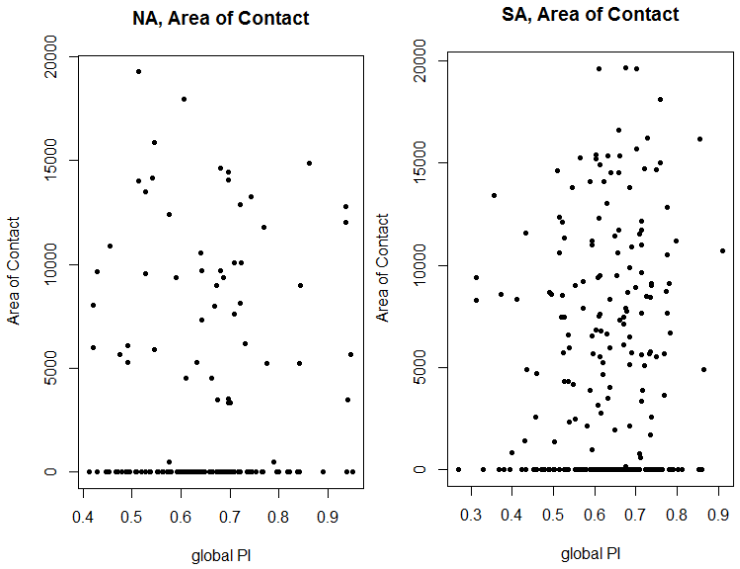


Figure 57: Relation between the typological PI and the area of contact. Left: North America, right: South America

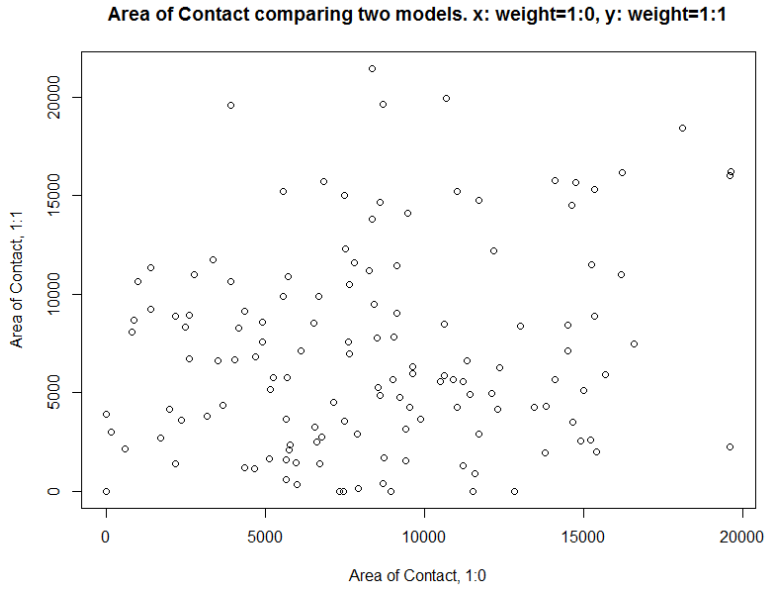


Figure 56: Relation between the AOC of two models. X-axis: b1 = 1, b2 = 0, y-axis: b1 = 1, b2 = 1

7 Discussion

The previous chapter has presented all important results of this thesis. Chapter 7 aims to discuss these results in order to provide further insights and to draw conclusions in Chapter 8. All results have been presented in a manner to answer the research questions in section 4.1. Therefore, the results are discussed in ascending order of research question number one to four, whereas the last two questions are feeding into the same discussion. As already shown in the previous chapter, the results and, consequently, the interpretation are related to spatial discrepancies whereas North and South America must be analyzed separately. Some controversial results within the spatial subsets as well as within the expected outcomes make it difficult to reach a clear conclusion. Nevertheless, the surprisingly distinguishing outcome highlights the complexity and spatial dependency of this phenomenon.

7.1 Scaling-issues of language particularity

RQ1: How do scaling-issues affect quantitative language particularity?

Hypothesis 1: Language particularity is highly dependent on the scaling-levels.

To answer this question, language particularity has been computed for different scaling-levels. The global PI includes all languages of the Americas: the focal splits the sampling in North and South America and the local PI measures similarity to the neighbors within 1000 km. The high impact of this sampling was shown in section 6.2. This insight will question a lot of studies and theories and, of course, it does draw attention to the use of this relative measurement. However, the aim is not to belittle previous investigations, but, above all, the idea is to create new methods to better understand language development. Consequently, the effect of different PI for different scaling levels should not be seen as a limitation, but rather, as a possibility for new interpretation. To achieve that, it is crucial to deal with the meaning of the changes of the particularity in different sampling methods.

In general, languages defined through a low particularity could be called *common.* However, it is important to be aware of the fact that they are only common in this specific sampling set. Consequently, a common language in a global scaling describes different properties of a language than common languages on a focal scaling. If the characteristics of all languages were randomly distributed over the whole Americas, there would not be any difference for the various particularity indices. Thus, the insight on the high impact of changing the subset allows to drawing conclusions

about the spatial distribution of language traits. The next section treats the reasons and meanings of the change of particularity for different scaling-levels.

Table 9 visualizes the main possibilities of change for two different scaling-levels. The particularity can stay either low or high, which is called *stable*. In addition, the change of the particularity can be *decreasing* or *increasing.* For each of these cases, an example is shown and discussed below. It was already pointed out that the values of the local PI are likely to be biased through the number of na's. Thus, this interpretation focuses on the meaning of the global and focal level. Nevertheless, possible interpretations of the local index are briefly discussed.

Table 9: Possible changes of the particularity for different levels

		Level 2	
		low	high
Level 1	low	stable	increasing
	high	decreasing	stable

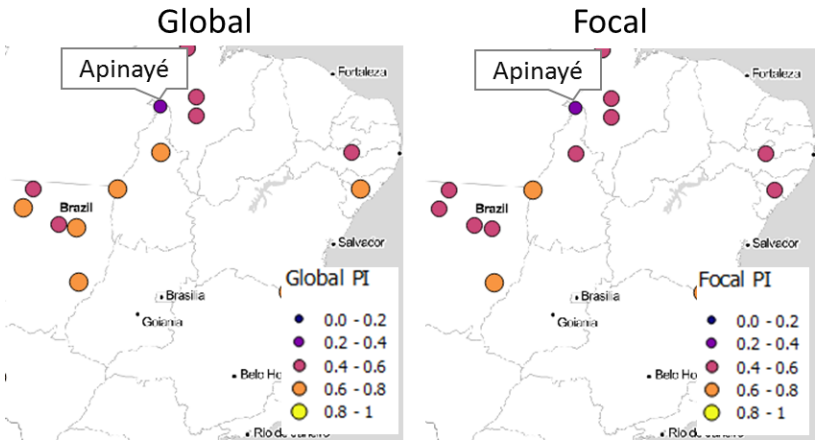


Figure 58: Example of stable development in global and focal TP: language Apinayé

The language *Apinayé* is an example of a stable and low value of particularity in two levels as shown in Figure 58. In both definitions, *Apinayé* represents a rather particular language in a global, as well as in a focal view.

The typological particularity is an average mean of different features. The feature values that are shared by most of the languages represent the typical properties of the sampling set. The value of the particularity does not allow defining which features have led to the result. So, in this example, it is not clear whether the same properties have defined the language as common. There are three main possibilities that have led to this classification in both scaling-levels. The first explanation could

be that the typical features of the Americas are the same as for North America. In this case, there would not be any difference between the indices at all. Nevertheless, other analyses have shown that this not true. More likely, there are some feature values that are the same for both continents. It could be possible that the language *Apinayé* mainly contains such characteristics that are common in both scaling-levels. Alternatively, it is possible that in the focal PI, different features have led coincidentally to a similar value as the global PI. This can be explained by an artificial language that contains ten features, where feature one to four are common for the Americas and six to ten for North America. Feature number five is neutral in both. Thus, while being in the global PI, the first four features have led to the low value, the main influence within the focal level lies within the features six to ten.

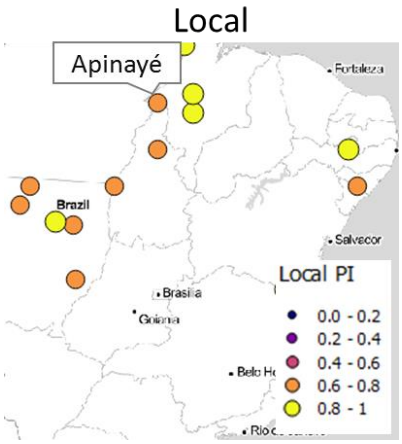


Figure 59: Local particularity of the language Apinayé.

The particularity in a local definition is different as shown in Figure 59, where *Apinayé* turns into a particular language, compared to its nearest neighbors. As we know, the language is typically American and South American, so, it is likely that this language point lies next to languages that are non-typical for either of them. Consequently, the typical features are not shared by its neighborhood and change from common to particular properties. In this case, the different levels of particularity can give more information on the spatial distribution of language characteristics than a general index would.

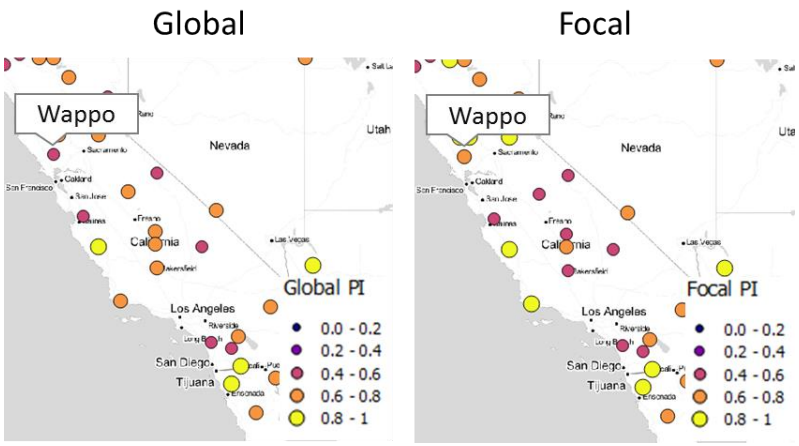


Figure 61: Example of increasing development in global and focal TP: language Wappo

The results have shown indications that in focal PI, languages in South America tend to be categorized as more common and languages in North America increase in their particularity. Figure 61 visualizes such a language. *Wappo* lies on the west coast of North America and is exemplary for an increasing value in smaller scaling. In other words, the language *Wappo* is common from a global point of view, but particular if only considering North America. Thus, it seems that a lot of the characteristics of *Wappo* have been shared in South America. As soon as the South American languages were excluded from the subset, these feature values turned into a minority and classified as particular.

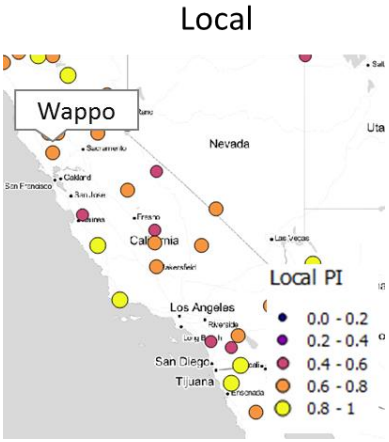


Figure 60: Local particularity of Wappo.

The local PI in Figure 60 underlies the assumption that has been made for the focal PI. That is, the language *Wappo* does not have typical North American features. As the language is located in a very

dense zone of language points, such a result would have been expected. It means that many other languages within this zone share feature values that are typical for North America.

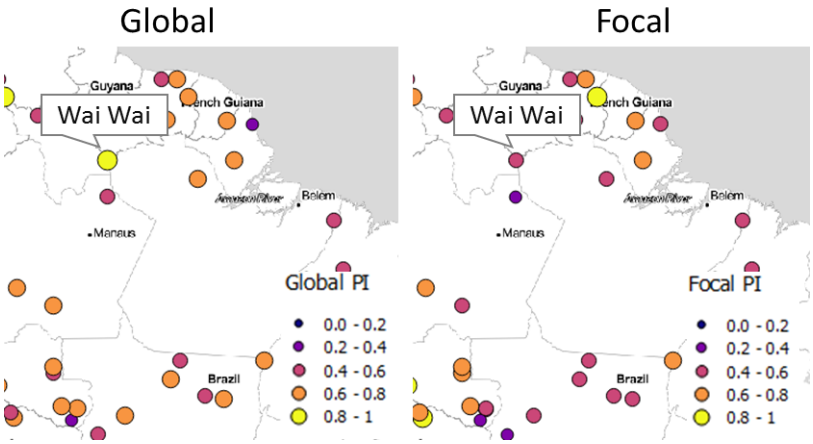


Figure 62: Example of decreasing development in global and focal TP: language Wai Wai

The language *Wai Wai* in Figure 62 shows the opposite of *Wappo* and has been defined as quite common when only compared to the south American languages, but very particular on a global scale. Hence, *Wai Wai* represents frequent characteristics of a South American language and may describe the main differences between the languages of the North to the South.

As already mentioned, interesting insights can be gained not only through the individual changing but also through the clustering of behavior. The existence of clusters denies the random distribution of spatial phenomena and shows the dependency of space on language typology. The most interesting areas of such clusters are visualized as zone *a*, *b* and *c* in Figure 63. The differences in global PI to focal PI are mainly negative. Thus, in general, there are a lot of languages that are classified as less particular.

Zone *a* shows that the particularity of most of the languages of South America is decreasing in case one does not consider the north. One reason might be that there are features that are very specific for the south. When one does not compare the languages to the upper continent, these few but significant features turn into normality. On the contrary, within the extent of *b*, many languages can be observed that share a positive change in their particularity. Some languages seem to have features that they have shared with the southern languages. By removing them from the sampling, they turn into rare features. It is not clear whether these are the same features that are responsible for the decrease of particularity in South America. However, one explanation could be that this language group has been in contact with the southern languages before they moved to South America.

The languages within the extent of *c* did only rarely change and represent a typology that is typical for the Americas and also for the northern subcontinent. Considering the history of migration from the Bering Strait to the south, these languages lie at a central point of the route and are probable to have been in contact with most of the communities (Nettle 1999). This could explain the shared features of the northern, as well as of the southern languages.

Central America has been categorized as part of the northern continent. The behavior of the languages close to South America represents a pattern of increasing values. In northern direction, clusters of decreasing values are found. For later analyses, it would be interesting how they would change if it were treated as a single focal zone.

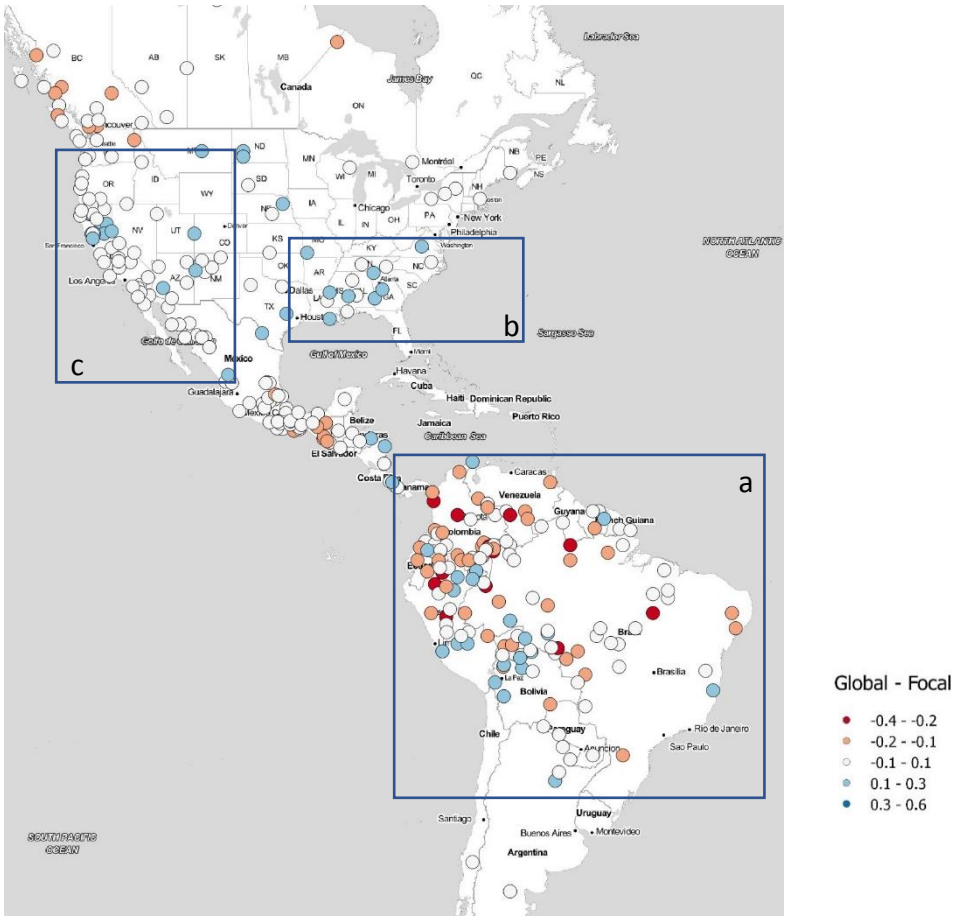


Figure 63: Impact of the parameter weight for the genealogical PI (top) and the global PI (bottom). To the left, the weighting of elevation to historical surface is 1:0, to the left 2:1.Figure 67: Different areas of interests in the change of PI

In combination with the global level, the expression of American average or normality has been used. We could say that the typology on a low level of scaling is the result of a very long time of social movements and developments that have created a baseline for all languages. To reduce the extent also limits the possible timeline in which these societies have been in contact. In other words, languages that are far away from each other only could have been in contact many thousand years ago. On the other hand, close languages represent a more recent contact situation. As a result of zooming out of a region, the processes go more and more back into the past. By zooming in, more recent developments can be observed.

The results and the discussion have shown that hypothesis 1 can be verified. Thus, the quantitative measurement of particularity of languages is highly dependent on the subset or scaling-level of the languages. In addition, this insight allows furthering the interpretation of spatial distribution and historical human movements.

7.2 IBD and Altitude as factors of Isolation

RQ2: How do simple factors of isolation explain language particularity in the Americas?

Hypothesis 2: Metric distance and altitude are simple factors for isolation and can explain language particularity in the Americas.

This research question addresses previous studies and assumptions in the linguistics (Nichols 2003, Holman et al. 2007). It intends to reproduce scientific insights for the very specific study area and measurement of language particularity. In a first step, the concept of IBD was analyzed. Thereby, indications for IBD could only be shown for genealogical particularity in South America. None of the typological PI did reflect IBD in neither of the areas. Due to the fact that the assumption that close languages are more similar than distant languages is well-established in linguistics. That is why the result of this analysis could be an indicator for the incorrectness of the measurement of the particularity. It is possible that the data set of WALS is not appropriate for such analysis based on the high number of na's. The methodologies have shown that the main limitations of the analysis of language traits are due to the dataset that is represented by the sparse information of WALS. There are some languages or features that have been determined in much more detail than others. In terms of quantitative analysis, features and languages that contain too little information must be dropped. As a result, the problems should be addressed that it is likely that the sparse features are

rather particular or rare and, consequently, this method leads to an underestimation of particular languages.

However, to date, the assumption of IBD was mainly done based on qualitative measurement and in the manner of genealogical characteristics. Thus, the missing correlation of metric distance to typological particularity could hint to a lack of insights in the field of linguistics. Also, most of the linguistic analyses have been conducted on a global scale. Only a few scientists have done quantitative investigations on the specific study area of the Americas. Hence, it is possible that the assumptions of IBD might be true from a global point of view and the Americas represent an exception or special case.

It is also important to point out the insight of the stronger correlation of genealogical particularity in South America compared to North America. Perhaps, the later colonization of South America could account for this fact. Due to the limited time frame, language families have had less time for dispersion. Languages with the same genealogical origin are closer to each other, because dispersion needs time, which these languages did not have. This questions whether genealogical particularity really measures language traits or rather historical patterns. In this case, the similarity of the languages would not be due to the accessibility or isolation of the languages, but it would be characterized by the initial position of the families (Bowern 2013).

In a second step, the suggested correlation of altitude to language particularity was determined. The analysis has shown that the correlation between genealogical particularity and altitude in this analysis is positive. This would mean that the higher the isolation of a language, the less particular it is. Especially in South America, many particular languages are located on low altitudes while common languages are found on high levels of altitude. The geography of South America and their historical population and civilizations could explain this phenomenon. The advanced civilization of the Inkas has been formative for the development in South America, which has spread over the highlands of the Andes mountains (Troll 1932). Thus, the similarity of the languages in the region of the Inkas at a high altitude is not surprising. This insight shows the difficulty of the definition of global theories of language development. Also, it shows that the amount of meters above sea is not a sufficient factor to describe isolation for areas that contain highlands such as South America.

Typological particularity did not show any clear relation to the altitude. The problem is the same as already explained in IBD. If the dataset of WALS cannot provide enough information, the index does not reflect language particularity and it is not possible to define insights of this result. Assuming that

the particularity indices are correct, the result implies that the factor of altitude is not appropriate to reflect isolation for the Americas.

In sum, this analysis has not been able to reproduce the insights from previous studies about the correlation of metric distance or altitude to language particularity. It was shown that both factors seem not to mirror isolation for the study area of the Americas.

7.3 Impact of reachability

The interpretation of RQ 3, if the model of reachability is appropriate in terms of language particularity, cannot be answered without the regard of RQ 4. Thus, the two research questions are examined together in the following section.

RQ3: What is an appropriate method to model accessibility in terms of language particularity?
Hypothesis 3: Archeology and GIS provide appropriate methods to model accessibility in term of language particularity.

RQ4: What is the impact of accessibility to language particularity?
Hypothesis 4: The level of accessibility due to environmental factors highly influences language particularity in a negative correlation.

The model of accumulative costs has been evaluated as the most suitable for the approach of this master thesis. Most importantly, it allows the analysis with a point of origin and without a point of destination. The principle of cost surfaces allows the inclusion of different data layers. In light of this, the most important layer in terms of human movement is the elevation model. In this analysis, the historical surface was considered, too. The model is flexible to add different costs and data sets. An important advantage of this model is that it allows defining a general degree of accessibility, independently of other language points and without pair-wise comparison. Besides the general determination of the accessibility of each language point, the model can easily be adapted. For instance, the interaction and spatial distribution of all other languages can be considered. Based on the reachability of the accumulative cost surfaces, the area of contact can be computed. Thus, with one model, different approaches and extensions can be performed.

Limitations

In general, a clear interpretation is rendered difficult due to many uncertainties within the model of reachability, as well as due to the provided data set of languages.

Input Data

The difficulty due to the sparse data set of the languages has been discussed in section 6.4.2. Besides the number of na's, there are additional limitations when it comes to modeling language particularity. Due to the lack of historical language databases, the analysis can only consider today's languages and their actual position. Moreover, each language is only represented by one point. Neither population size nor density nor the area where the language is spoken, can be included. The dataset reflects language characteristics and their locations of today.

The method to determine language particularity assumes all features to have the same probability and to be transformed due to either isolation or contact. As already pointed out, the available language data represents the recent situation. Due to the slow rate of changes of languages, it would be reasonable to model accessibility of the past. Because this thesis will not model the past locations of the languages, it assumes that today's languages are still representative to gain insights about the development through a historical environment. The analysis includes the historical land surface. Importantly, there is lack of control and verification because of the fact that this information is the product of a model.

Model

There are still a lot uncertainties in terms of the parameter of cost transformation, weighting and threshold. Even for commonly used layers, such as the elevation model, the determination of the cost transformation is not clear. The more layers are added to the model, the more complicated the definition of all parameters gets. Especially the determination of the weight of each layer would demand a lot of qualitative studies. Figure 64 visualizes the impact of the weighting on two examples in North America. While the genealogical PI tends to a negative regression when considering the elevation model alone, the regression turns to positive as soon as the historical surface is added. The regression of the global PI to reachability lower plots becomes stronger when adding the historical surface. It is important to be aware of the uncertainty in the cost assignment of the historical surface that wields major influence on the outcome.

Due to missing computer power, only two information layers could be determined in detail. The limitation of the computer power also leads to a limited outcome of results. The parameters cannot be determined in innumerable analyses. In these analyses, the model has been fitted through some parameters and the best suitable model for each region and particularity index has been defined. It is possible that there are models that are more suitable, but which have not been evaluated due to time limitations.

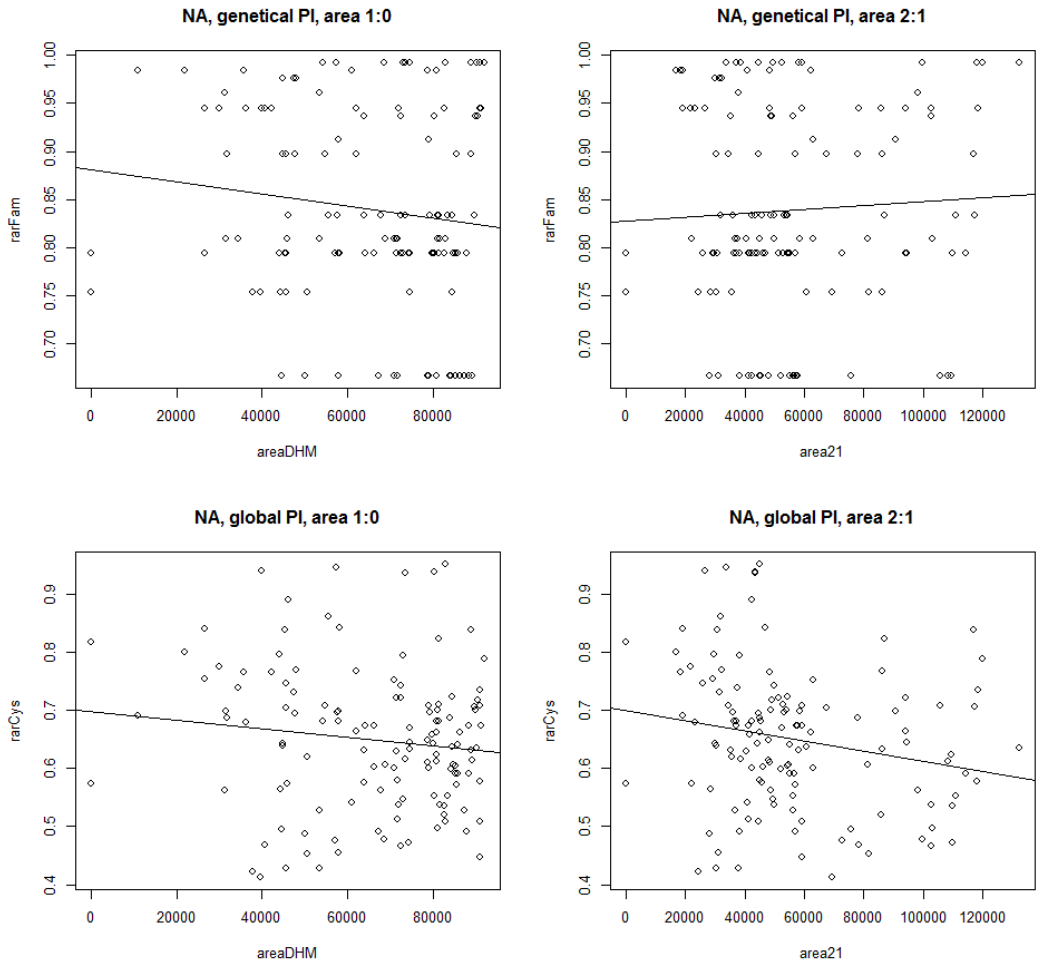


Figure 64: Impact of the parameter weight for the genealogical PI (top) and the global PI (bottom). To the left, the weighting of elevation to historical surface is 1:0, to the left 2:1.

As already said, the model allows the inclusion of many different data layers. However, the access to many important factors of human movement, as for instance socioeconomic dominance or religion barriers, is very limited. Further data fitting to this study area and in quantitative form could not be found.

Figure 65 summarizes all limitations and uncertainties of this analysis that have been discussed in this Chapter. The sparse information of the linguistic dataset and their simplification of a single point, as well as the unknown parameters and missing data layers all present sources of error.

In total, language particularity could not be modeled through accumulative cost surfaces. The high sensitivity of the parameters and the lack of scientific research complicate definitions of clear outcomes. However, it does neither mean that the relation between language particularity and isolation is inexistent, nor that the approach to the model is inappropriate. Even if the statistical values are not outstanding, some important insights can be found in this analysis.

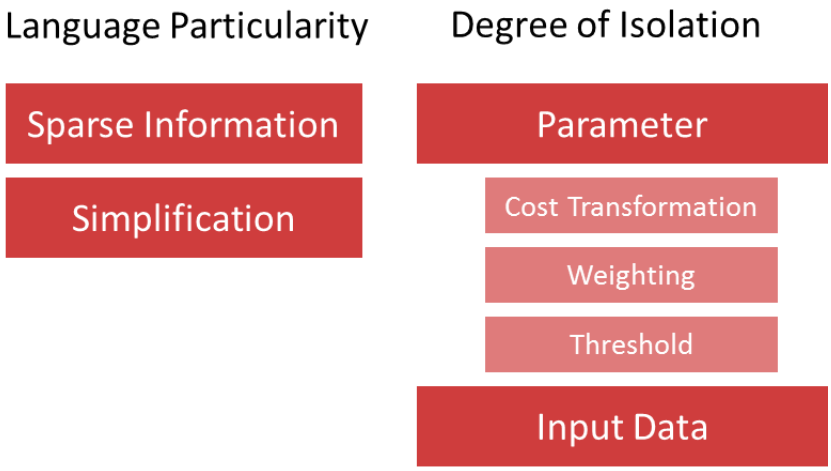


Figure 65: Limitations and uncertainties for the variable of language particularity as well as the degree of isolation

Genealogical particularity

The first approach to define language particularity is based on family membership. Weak relation to the modeled reachability in both subcontinents was found when only considering the elevation model. Historical land surface could never enhance a correlation. Interestingly, at least some languages in North America show the expected correlation. To clarify, large families that are defined as less particular, are located in well accessible regions. On the contrary, genealogical particular languages in South America tend to be located in well accessible areas. The same unexpected result for genealogical particularity in South America was found for the relation to the factor of altitude. The explanation of the dispersion of the Inkas within mountainous areas can be adapted for this result. It may be possible that the language group in the upper right corner in Figure 66 has been characterized through the population of the Inkas. Thus, the area of reachability is low due to the mountainous terrain but the genealogical PI is high due to the history of the Inkas.

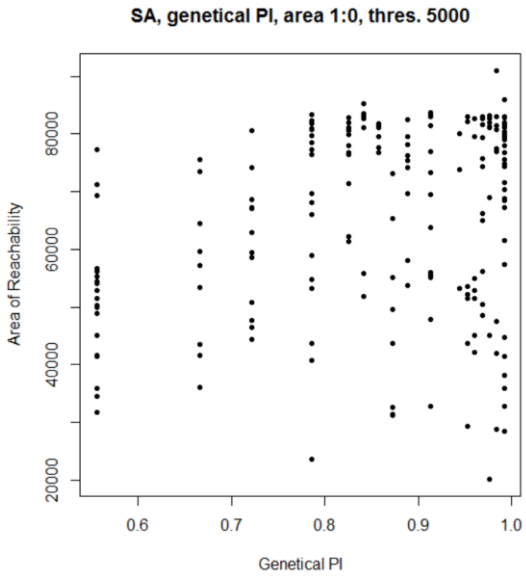


Figure 66: Relation between the genealogical PI and the area of reachability in South America.

Typological Particularity

Differently to the genealogical PI, typological particularity could best be explained through the inclusion of the historical land surface. In North America, the historical land surface has even been weighted more than the elevation model. The different weighting of the historical land surface leads to a discussion on the distribution of the land surfaces. As shown in Figure 67, the distribution of some classes varies with regards to a comparison of the two subcontinents. It may be possible that

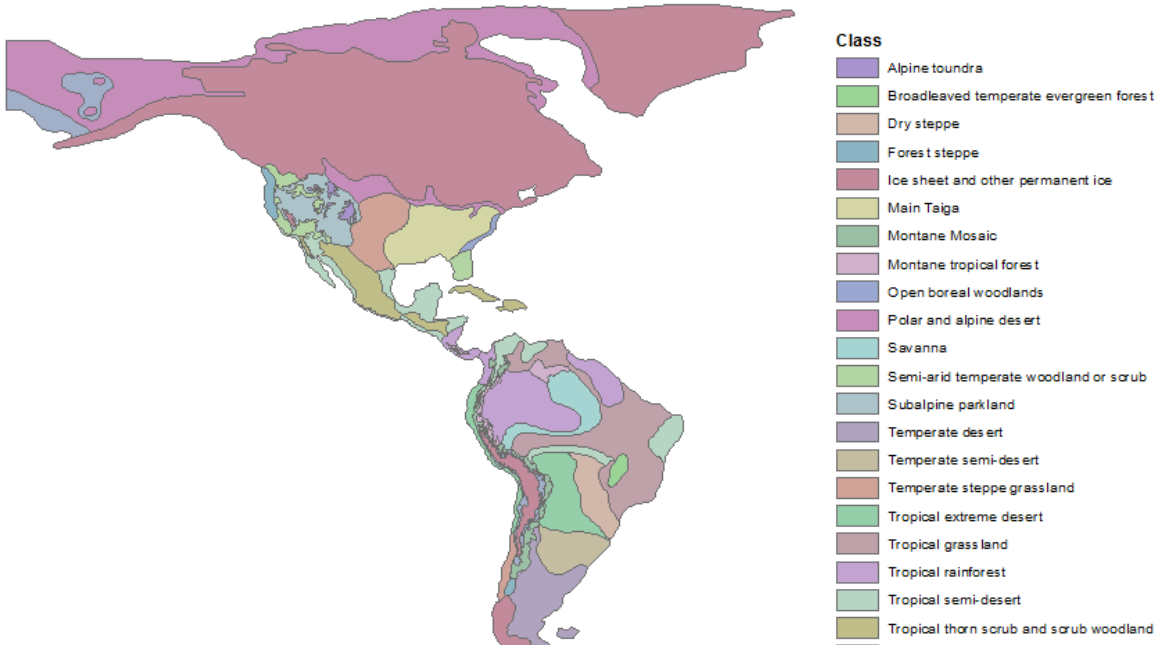


Figure 67: Visualization of the historical surface of the Americas

the influence of different classes is higher than others. For instance, the historical land surface seems more important in North America. As this part of the continent is dominated by the surface *ice-sheet and other permanent ice,* this surface could have a higher influence on language development as others. Another risk is the misclassification of costs of some surfaces. Namely, if a surface that mainly appears in South America has been wrongly assigned, the model of the south would be less realistic than the northern model.

It was shown that the reachability of the languages is clustered in two in North America. This behavior has been underlined in Figure 68. Indeed, the first peak of low reachability is most obvious. The second language group can be recognized by a higher reachability. In between, there are only some single language points. To put it in drastic wording, this would mean that this model classifies languages in only two groups that are *well* or *hard reachable.* From this follows that instead of explaining languages according to this factor, this could be used as a classification. Based on this classification, investigations due to different factors could be conducted.

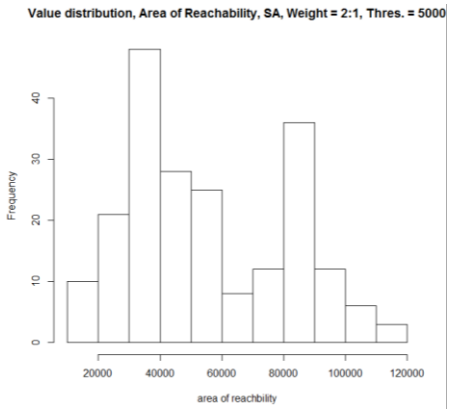


Figure 68: Value distribution of the area of reachability in North America.

In addition, the area of reachability does explain best global PI in North America and focal PI in South America. A vague interpretation could be based on the time of migration. Considering that all communities have arrived in North America, all language traits have been shaped through the historical land surface and elevation of the North. Only some of them have moved southwards through the years. Thus, the language traits that can be explained through the model of South America only reflect the languages within focal scale-level. As already said above, the assumption that the different levels of scaling reflect the component of time must be treated carefully.

Area of Contact

It has been argued that reachability reflects language particularity, because the more a community is isolated, the smaller the probability to get in contact with others. The model of reachability has only considered the environment, but ignored the distribution of the points. Nevertheless, it is possible that many points lie close to each other within an area of low reachability. Thus, the probability of contact is also due to the density of the language points. This is taken into account in the advanced model that computes the area of contact.

As this analysis is an extension of the reachability analysis, only the models that have been computed in the upper part could be evaluated. Unfortunately, also the computation of the areas of contact is very costly. However, for all models that have been tested, the correlation between language particularity and the area of contact is weak. Relations to typological particularity could neither be observed for North nor for South America.

The clearest relation could be shown for the genealogical particularity in South America. This is very interesting when considering the correlation of this index to the area of reachability in Figure 69 where the result shows the opposite of the expected results. Conversely, the area of contact reflects the expected outcome. For example, most of the languages with an area of contact of zero tend to higher particularity. In Figure 66, the language group in the upper corner was interpreted as languages that share similarity due to the population of the Inkas. Figure 69 shows that some of these languages, despite their large areas of reachability, have very small areas of contact. In this example, the area of contact seems more reasonable as the area of reachability.

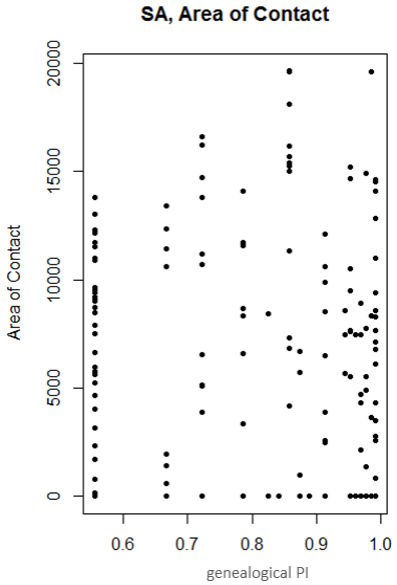


Figure 69: Relation between genealogical PI and the AOC

Nevertheless, the area of contact and its benefits could not persuade enough to explain language particularity. It was shown that the computed areas of reachability show a high amount of zero values, which could be changed by using a higher threshold. Unfortunately, this could not be evaluated within this work due to the limitations of computer power.

Another reason for the missing correlation could lie within the time stamp of the language points. It is quite probable to assume movement of the societies in the last years. The reachability polygons might reflect a general value of isolation within the region of a language point. The area of contact might be more dependent on the exact location of the language point. This means that while a movement of the language points of 500 km might result in a similar value of reachability, the area of contact would be highly influenced. Thus, the area of contact is not able to model the historical situation and consequently does not reflect language development.
However, a clear interpretation would require a more detailed analysis by many more parameters. In this case, the model could neither verify nor falsify the hypothesis that language particularity is influenced through the contact between communities.

Overall, different methods from the field of archeology and GIS have been combined and a model of accumulative costs has been developed that defines the reachability of language points. In terms of human movements, it allows the addition of many layers and the flexibility to weight them. As a result, human movement can be modeled more realistically and appropriately to get insights on language development and its dependency to the environment. In practice, the lack of scientific research and the missing computer power led to many uncertainties of some parameters. Thus, this thesis has generated a basic framework of a model of reachability that has potential to be adapted and refined. With this in mind, the hypothesis 3 is verified.
The accumulative model is not able to explain language particularity in the Americas and the hypothesis 4 could not be verified. Due to the different source of errors in the datasets and the parameters, it does not mean that the assumption of higher particularity in isolated regions is wrong. This analysis has shown that the development of languages is a highly complex phenomenon that is strongly dependent on the geography. The results have been able to reflect some of these spatial dependencies and shown that the cooperation of GIS provides new possibilities in linguistic studies.

8 Conclusion

This thesis has strengthened the cooperation between the fields of linguistics, archeology and GIS. Previous investigations into this cooperation have determined a strong connection between language development and geography. Especially the factor of geographical isolation has shown high impact. However, to date, the concept and implementation of geographic isolation are mainly kept very simple. To address this lack of research, this thesis has presented a model as a more realistic approach to defining isolation, and reachability in the manner of historical human movement respectively. With the help of this model, the impact of geographical reachability to language particularity has been determined. The basics of the model lie within an accumulative cost surface that allows the inclusion of different information layers. Within this work, the model has been fed with the elevation model and the historical land surface to explain language particularity. The limited input data allows to simply explain the function of the model and to properly explain the impact of the parameters. Unfortunately, the lack of scientific insights on these parameters – weighting, cost assignment and threshold - leads to limited interpretation. Another difficulty lies within the data set of the languages. Due to the sparse information of the data set, the measurement of typological particularity is likely to be biased by non-available data.

Overall, the model was not able to explain language particularity in the Americas. However, there are important insights in terms of language particularity as well as in terms of the concept and model of reachability. With the help of different scaling-level, this thesis has shown that the measurement of language particularity is always relative to a specific sampling set or spatial extent. Consequently, a language might be classified as very particular on a low zoom level, but represents a common language for a high zoom level. To date, scientific research has barely brought attention to this fact. As a result, the use of different scaling-levels can offer new insights into the spatial distribution of language properties. In addition, it was shown that the definition of language particularity due to genealogical or typological properties is crucial. This analysis could not find any clear correlation between the two measurements. This result draws attention to definitions and terms of language particularity.

Furthermore, this thesis has shown that the definition of isolation and reachability for human movement respectively represents a highly complex process that still commands a lot of potentials. First of all, the concept of reachability can be defined through different components as the elevation model, the land surface, religion or socioeconomic situations.

All these factors must then be parameterized to form a model. This work has shown that the implementation of only two factors did already transcend the scientific insights into human movement. Thus, the lack of knowledge about human movements prevented the exact definition of the degree of isolation for languages. However, this work has provided a basic framework to model reachability and that shows potential that is to be redefined.

Due to these uncertainties, clear conclusions on the behavior of language development due to geography could not be drawn. An important insight of this analysis is that there are languages that could be explained by the model, while there are languages that show total controversial dependencies. This insight highlights the existence of geographical dependencies that are not determined enough yet. It was also shown that the complexity and history of the society must always be taken into account to explain facts and exceptions.

This analysis has shown that the development of languages is a highly complex phenomenon that is strongly dependent on the geography. The conclusions drawn from the results have been able to reflect on some of these spatial dependencies and have shown that the cooperation of GIS provides new possibilities in linguistic studies.

9 Outlook

The use of GIS in quantitative linguistics offers a large potential for further research. Considering the particular analysis of this thesis, all the dataset of languages must be completed in a much higher detail to begin with. As the collection of linguistic data is mostly done in qualitative studies, the development of appropriate statistical measurements should be addressed more extensively.

In light of the accumulative cost model, there is a clear need for more investigation of the unknown parameters. To date, most of the human movements are modeled by the single use of the distance and elevation. There is the need for further investigation on different information layer such as land surface, for instance. Importantly, the focus should not only lie within the conceptual information basis, but also within the actual implementation thereof. Empirical and computational studies could show the exact cost assignment of environmental components as well as the influence of different data layers. With the help of such studies, it would be easier to define parameters as the weighting or cost transformation.

This thesis has developed a method to compute geographical reachability. However, in many cases the method was too costly in terms of computational power. Thus, there might be similar approaches or adaptations of the model that ensure higher efficiency.

Abbreviation

Abbreviation	Explanation
AOC	Area of Contact
BB	Bounding Box
GIS	Geographic Information Science/System
LCP	Least-cost-path
NA	North America
na	Non-available information
IBD	Isolated-by-distance
Pmi	Point-wise mutual information
PI	Particularity index / indices
SA	South America
WALS	World Atlas of Language Structures

Bibliography

Aikhenvald, A. Y. (2003). Mechanisms of change in areal diffusion : new morphology and language contact. *Linguistics*, *39*, 1–29.

Aikhenvald, A. Y. (1992). *Grammars in Contact: A Crosslinguistic Typology*. (A. Y. Aikhenvald & R. W. Dixon, Eds.). New York: Oxford University Press Inc.

Aitchison, J. (2005). Language change: progress or decay? In *The Routledge Companion to Semiotics and Linguistics* (3rd ed., pp. 111–120). Cambridge University Press Press.

Anderson, D., & Gillam, J. (2000). Paleoindian colonization of the Americas: implications from an examination of physiography, demography, and artifact distribution. *American Antiquity*, *61*(1), 43–66.

Bowern, C. (2013). Relatedness as a Factor in Language Contact. *Journal of Language Contact*, *6*, 411–432.

Cysouw, M. (2011). Quantitative explorations of the worldwide distribution of rare characteristics, or: the exceptionality of northwestern European languages. In H. J. Simon & H. Wiese (Eds.), *Expecting the Unexpected: Expections in Grammar* (pp. 411–431). De Gruyter Mouton.

Cysouw, M., & Comrie, B. (2009). How varied typologically are the languages in Africa? In R. Botha & C. Knight (Eds.), *The Cradle of Language* (12th ed., pp. 208–222). OUP Oxford.

Cysouw, M. (2018). qlcMatrix. Retrieved June 20, from https://github.com/cysouw/qlcMatrix

Dahl, Ö. (2008). An exercise in a posteriori language sampling. *Östen Dahl (Stockholm)*, *61*(3), 208–220.

Dahl, Ö., Gillam, C., Anderson, D., Iriarte, J., & Copé, S. (2011). Linguistic Diversity Zones and Cartographic Moedling: GIS as a Method for Understanding the Prehistory of Lowland South America. In A. Hornborg & J. Hill (Eds.), *Ethnicity in Ancient Amazonia: Reconstructing Past Identities from Archeology, Linguistics, and Ethnohistory* (pp. 211–224). Colorado: University Press of Colorado.

Daumé Iii, H. (2009). Non-Parametric Bayesian Areal Linguistics. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 593–601).

Dixon, E. J. (2011). Human colonization of the Americas: timing, technology and process. *Quaternary Science Review*, *20*(1–3), 277–299.

Dryer, M. S., & Haspelmath, M. (2013). The World Atlas of Language Structures Online. Retrieved June 4, 2018, from http://wals.info/

Georgi, R., Xia, F., & Lewis, W. (2010). Comparing language similarity across genetic and typologically-based groupings. *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics.

Geurs, K. T., & van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography*, *12*(2), 127–140.

Gimond, M. (2017). Intro to GIS and Spatial Analysis. Retrieved June 4, 2018, from https://mgimond.github.io/Spatial/index.html

Greenberg, J. H. (1987). Language in the Americas. Stanford, California: Stanford University Press.

Gruhn, R. (1988). Linguistic Evidence in Support of the Coastal Route of Earliest Entry Into the New World. *Man*, *23*(1), 77.

Heggarty, P. (2015). Prehistory through language and archaeology. In C. Bowern & B. Evans (Eds.), *The Routledge handbook of historical linguistics* (1st ed., pp. 598–628). New York, New York, USA: Routledge.

Holman, E. W., Schulze, C., Stauffer, D., & Wichmann, S. (2007). On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguistic Typology*, *11*(2), 393–421.

Howey, M. C. L. (2007). Using multi-criteria cost surface analysis to explore past regional landscapes: a case study of ritual activity and social interaction in Michigan, AD 1200–1600. *Journal of Archaeological Science*, *34*(11), 1830–1846.

Howey, M. C. L. (2011). Multiple pathways across past landscapes: circuit theory as a complementary geospatial method to least cost path for modeling past movement. *Journal of Archaeological Science*, *38*(10), 2523–2535.

Llobera, M., Fábrega-Álvarez, P., & Parcero-Oubiña, C. (2011). Order in movement: a GIS approach to accessibility. *Journal of Archaeological Science*, *38*(4), 843–851.

Lucas, C. (2015). Contact-induced language change. In C. Bowern & E. Bethwyn (Eds.), *The Routledge Handbook of Historical Linguistics* (pp. 519–536). London: Routledge.

McGregor, W. (2015). *Linguistics : an introduction* (2nd ed.). Bloomsbury.

Mckillop, H. (2005). Finds in Belize document Late Classic Maya salt making and canoe transport. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(15), 5630–5634.

Murrieta-Flores, P. (2012). Understanding human movement through spatial technologies. The role of natural areas of transit in the Late Prehistory of South-western Iberia. *Trabajos de Prehistoria*, *69*(1), 103–122.

Murrieta-Flores, P. A. (2009). Traveling in a Prehistoric Landscape: Exploring the Influences that Shaped Human Movement. In B. Frischer, J. Webb Crawford, & D. Koller (Eds.), *Making history interactive: Computer applications and quantitative methods in archaeology (CAA)* (pp. 258–276).

Nettle, D. (1999). Linguistic diversity of the Americas can be reconciled with a recent colonization. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(6), 3325–9.

Nichols, J. (1990). Linguistic Diversity and the First Settlement of the New World. *Language*, *66*(3), 475.

Nichols, J. (1992). Linguistic Diversity in Space and Time. University of Chicago Press.

Nichols, J. (2013). The vertical archipelage: Adding the third dimension to linguistic geography. In P. Auer, M. Hilpert, A. Stukenbrock, & B. Szmrecsanyi (Eds.), *Space in language and linguistics : geographical, interactional, and cognitive perspectives* (pp. 38–60). De Gruyter.

Perry, B., & Gesler, W. (2000). Physical access to primary health care in Andean Bolivia. *Social Science & Medicine*, *50*(9), 1177–1188.

Ranacher, P., von Gijn, R., & Derungs, C. (2017). *Identifying probable pathways of language diffusion in South America*. Wageningen.

Ray, N., & Adams, J. M. (2001). A GIS-based Vegetation Map of the World at the Last Glacial Maximum (25,000-15,000 BP). *Internet Archaeology*, (11).

Rees, W. G. (2004). Least-cost paths in mountainous terrain. *Computers & Geosciences*, *30*(3), 203–209.

Rogers, R. A., Martin, L. D., & Nicklas, T. D. (1990). Special Paper: Ice-Age Geography and the Distribution of Native North American Languages. *Journal of Biogeography*, *17*(2), 131.

Sankoff, G. (2002). Linguistic Outcomes of Language Contact. In P. Fletcher, B. Macwhinney, W. J. Hardcastle, J. Laver, A. Spencer, A. Zwicky, … G. Ward (Eds.), *The Handbook of Language Variation and Change Edited* (pp. 638–668). Blackwell Publishing Ltd.

Sokal, R. R., & Wartenberg, D. E. (1983). A Test of Spatial Autocorrection Analysis using and Isolation-By-Distance Model. *Genetics*, *105*(1), 219–237.

Soule, R. G., & Goldman, R. F. (1972). Terrain coefficients for energy cost prediction, *32*(5).

Thomason, S. G. (2001). *Language Contact*. Edinburgh: Edinburgh University Press.

Thomason, S. G. (2008). Contact as a Source of Language Change. In *The Handbook of Historical Linguistics* (pp. 687–712). Oxford, UK: Blackwell Publishing Ltd.

Tobler, W. R. (1970). A computer movie simulating urban growth inthe Detroit region. *Economic Geography*, *46*, 234–240

Tobler, W. R. (1993). Three Presentations on Geographical Analysis and Modeling. *Speculations on the Geometry of Geography; and Global Spatial Analysis*, *93*(1), 1–93.

Todd Jobe, R., White, P. S., & Todd, R. (2009). A New Cost-Distance Model for Human Accessibility and an Evaluation of Accessibility Bias in Permanent Vegetation Plots in Great Smoky Mountains National Park, USA. *Journal of Vegetation Science Journal of Vegetation Science*, *20*(6), 1099–1109.

USGS (2015). USGS - Science for a changing World. Retrieved June 14, 2018, from https://lta.cr.usgs.gov/get_data

Van Etten, J. (2017). *R Package gdistance: Distances and Routes on Geographical Grids*.

White, D. A., & Barber, S. B. (2012). Geospatial modeling of pedestrian transportation networks: a case study from precolumbian Oaxaca, Mexico. *Journal of Archaeological Science*, *39*(8), 2684–2696.

Wichmann, S., Holmann, E., Stauffer, D., & Brown, C. H. (2011). Similarities among languages of the Americas: An exploration of the WALS evidence. *Journal of Language Relationship*, *67*(5), 1–163.

Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Place, Date _____ Hella Mönkeberg _____