Bioinformatic methods in calculating language relationships

Anke Lüdeling and Ulf Leser

ESSLLI 2008
20th European Summer School in Logic, Language and Information
4–15 August 2008
Freie und Hansestadt Hamburg, Germany

Programme Committee. Enrico Franconi (Bolzano, Italy), Petra Hendriks (Groningen, The Netherlands), Michael Kaminski (Haifa, Israel), Benedikt Löwe (Amsterdam, The Netherlands & Hamburg, Germany) Massimo Poesio (Colchester, United Kingdom), Philippe Schlenker (Los Angeles CA, United States of America), Khalil Sima'an (Amsterdam, The Netherlands), Rineke Verbrugge (**Chair**, Groningen, The Netherlands).

Organizing Committee. Stefan Bold (Bonn, Germany), Hannah König (Hamburg, Germany), Benedikt Löwe (**chair**, Amsterdam, The Netherlands & Hamburg, Germany), Sanchit Saraf (Kanpur, India), Sara Uckelman (Amsterdam, The Netherlands), Hans van Ditmarsch (**chair**, Otago, New Zealand & Toulouse, France), Peter van Ormondt (Amsterdam, The Netherlands).

`http://www.illc.uva.nl/ESSLLI2008/`
`esslli2008@science.uva.nl`

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

ESSLLI 2008 is organized by the Universität Hamburg under the auspices of the *Association for Logic, Language and Information* (FoLLI). The *Institute for Logic, Language and Computation* (ILLC) of the *Universiteit van Amsterdam* is providing important infrastructural support. Within the Universität Hamburg, ESSLLI 2008 is sponsored by the Departments *Informatik*, *Mathematik*, *Philosophie*, and *Sprache, Literatur, Medien I*, the *Fakultät für Mathematik, Informatik und Naturwissenschaften*, the *Zentrum für Sprachwissenschaft*, and the *Regionales Rechenzentrum*. ESSLLI 2008 is an event of the *Jahr der Mathematik 2008*. Further sponsors include the *Deutsche Forschungsgemeinschaft* (DFG), the Marie Curie Research Training Site GLoRiClass, the European Chapter of the Association for Computational Linguistics, the *Hamburgische Wissenschaftliche Stiftung*, the Kurt Gödel Society, Sun Microsystems, the Association for Symbolic Logic (ASL), and the European Association for Theoretical Computer Science (EATCS). The official airline of ESSLLI 2008 is Lufthansa; the book prize of the student session is sponsored by *Springer Verlag*.

Anke Lüdeling and Ulf Leser

Bioinformatic methods in calculating language relationships

Course Material. 20th European Summer School in Logic, Language and Information (ESSLLI 2008), Freie und Hansestadt Hamburg, Germany, 4–15 August 2008

Simulating and reconstructing language change

Mirko Hochmuth, Anke Lüdeling, Ulf Leser

Humboldt-Universität zu Berlin, Institut für Informatik
{mirko.hochmuth, leser}@informatik.hu-berlin.de
Anke.Luedeling@rz.hu-berlin.de

In this work we probe phylogenetic algorithms for their ability to reconstruct historic language relationships. We present a formal model for the development of languages incorporating vertical (genealogical) and horizontal (language contact) effects. As a distinctive feature, we also added a geographic model to mimic the effects of constrained population movements. Using our model, we generated a large number of simulated language histories whose results were analyzed by a variety of established phylogenetic algorithms. Therein, we systematically investigated the effects of different contact intensities and of geographic as well es genealogic topologies. We found that tree-based algorithms are robust under a variety of different settings and are capable of inferring (parts of) the relationships correctly even under high levels of network-like influences. We also studied the SplitsTree algorithm which should be more appropriate to cope with network-like effects. However, although SplitsTree clearly performs better in some settings, it generally shows a rather erratic behavior.

Keywords: language change simulation, phylogenetic network, computational linguistics

1 Introduction

Languages are created and go extinct in processes that are often compared to the evolution of biological species [34]. This is due to the fact that both languages and species undergo similar effects. Biological species evolve over time due to the interplay of random mutations and selection. Language change is often explained in a variationist model where several similar variant exist next to each other and one of them slowly wins over the other (there are many different view on how the evolutionary metaphor can be applied to languages, for recent approaches see e.g. [5, 23, 31]). In biology, small selected changes in the genetic code of a species accumulate over time and may eventually lead to a change in the phenotype. In the most extreme case, new species emerge when sub-populations migrate into a new area with different environmental processes; similarly, new languages may emerge when groups of people leave their homeland and settle in distant areas.

Although many of these similarities are not as convincing as they appear at first sight (we will give a more thorough discussion on this issue in Chapter 2), they led to the idea of computing language relationships by applying methods from biology for determining the relationships between species, in particular phylogenetic algorithms, see eg. [10]. Since the advent of modern molecular biology, relationships between species are determined based on their DNA [8]. The underlying idea is – described in a very simplistic manner – that all species originate from the same root in a tree-like fashion. Thus, there once was a DNA sequence from which the sequences of all species originate. If we now chose a gene that presumably was already present in this root sequence and whose function is important still [1], then all living species should have a copy of this gene. If we consider two fixed species, then their copies of this gene will be the more different the longer ago these species split, as both copies had more time to undergo independent evolution and will thus have accumulated different mutations. Phylogenetic algorithms solve the following problem: Given a sequences of the "same" gene in different species, reconstruct the (best, most likely, most parsimonious ...) tree of speciation events connecting these species. This tree is called a gene tree. Usually, several gene trees are combined to infer a species tree [29].

If we adopt this idea to languages, we derive at a model as follows. New languages are born as a language community splits up and each of the new communities develops in its own distinctive way. These splits most often occur due to geographic separation, but also social or cultural differences can lead to a separation. Over time, the differences between the languages of the different communities become bigger and at some point in time two speakers of these languages do not understand each other anymore. Thus, first dialects (or sociolects), and later distinct languages originate. To apply a phylogenetic algorithm, one may choose a particular concept whose meaning is present in all languages under study and compute a word tree from the different words used for this concept in the different languages; finally, by averaging over different word trees, a language tree can be computed. However, at which representation of a word this computation should be applied, is not as clear as one might expect. Usually, a phonological representation is used, but this approach has the problem that the way how words have been pronounced in earlier language stages often is not clear.

One fundamental assumption we made until now is that species and languages (which will from now on both be called taxa) evolve in a tree-like fashion. This implies that every taxon has exactly one parent taxon [2]. Any changes in its DNA or language wrt. to the parents DNA or language are a result of evolutionary processes on the DNA or language itself, and other taxa have no influence any more. Using this assumption, many algorithms have been developed to reconstruct the taxa tree [26]. Among the most commonly used are the methods

[1] For instance, one can use genes that are vital for central parts of all species, such as DNA copying or synthesis of amino acids.

[2] Note that this assumption clearly is false for the development of individuals in many species; for instance, humans have two parents.

maximum parsimony and *neighbor joining* [8]. In biology, these algorithms have proven their reliability in a many simulated and real studies [14]. They have also been applied numerous times for computing the relationship between languages [12, 25, 28, 30].

However, it is very questionable whether the tree assumption of taxa is valid for languages. Various effects are known where languages change due to their contact to other languages [3], i.e., where taxa influence each other in a manner that is not genealogical. If two separated language communities come in contact with each other, linguistic characteristics can be exchanged between them. Furthermore, two or more languages may melt to form pidgin or creole languages. These two processes hurt the fundamental assumption underlying all algorithms for phylogenetic trees. This does not imply that tree-based algorithms might not also be useful in this case. First, they can be applied to a carefully chosen set of properties of a language (such as a concept list [36]) or a set of syntactical properties [12]. However, this method always requires a very much debatable step of chosing such properties. Second, the tree-based algorithms are somewhat robust to noise, and may still successfully infer evolutionary relationships despite non-tree signals in the data. Especially when language contact is sparse, the deviations from the tree model might very well be out-weighted by the majority of words following the tree-like evolution, and the the true genealogical relationships might still be found b those algorithms despite the noise.

In this report, we study exactly this question. We report the preliminary results of a study where we quantitatively simulate language evolution and apply several phylogenetic algorithms to the result. By comparing the reconstructed phylogeny with the true one (logged during the simulation), we may judge the capability of different phylogenetic algorithms to cope with non-tree effects in evolution. We studied two tree-based and one network-based algorithm. Our simulation is based on a simple model of language evolution encompassing effects of language creation and extinction, borrowing of words, and isolated and systematic changes in the pronunciation of words. As lexicon we choose the Swadesh 200 list [36]. All comparisons of words are based on phonological transcriptions. For our experiment we follow the long-standing tradition in historical linguistics tomainly look at phonological change. We are aware of the fact that in order to describe language change one would also consider changes on other linguistic levels.

The results of our study are promising, but also point to various open questions. First, we observe that no agreed-upon model for describing network-like phylogenetic relationships exist. For the most suitable model, no publicly available implementation of a reconstruction algorithm exists, which hinders us from evaluating it. We also find several interesting effects of the choice of distance measure between phylogenetic graphs which are not resolved yet. Despite these shortcomings, our study shows that all evaluated algorithms are capable of detecting phylogenetic signals even in cases of extreme noise, and that, for a large class of

[3] Such effects, known as *horizontal gene transfer* and *hybridization*, also exist in biology, but almost exclusively in the realms of bacteria and of plants.

settings, the network-based algorithm outperforms the tree-based methods. Surprisingly, we found that this is not true for settings with very little contact, a behavior which we can explain by the infamous long-branch attraction that is well known from tree-like models. Overall, we believe that such simulation studies are a valuable method to judge the ability of algorithms to come closer to the true processes underlying language evolution than possible with current methods.

The paper is structured as follows. In the next chapter, we give a more detailed discussion on the comparability of biological evolution and language evolution. Chapter 3 describes the most important effects of language change and how we included them into our model for language change, which is explained in Chapter 4. In Chapter 5 we describe the algorithms used to infer the phylogeny of the simulated languages. Chapter 6 explains the reconstruction setup and the evaluation of the inferred graphs. In Chapter 7, we present the results of various experiments which are discussed in Chapter 8. We conclude our work with a summary and an outlook on future work.

Notes

In this work, different types of markup are used to distinguish characters from the *international phonetic alphabet* (IPA) and from *phonemes*. Single square brackets mark IPA–characters or phonetic spellings of whole words. Double square brackets mark the notation in which a word is represented in our implementation. This representation differs partially from the phonetic spelling due to technical reasons (see Section 6). Slashes mark phonemes. For example, [ç][4] marks the sound of the phoneme /ç/, the phonetic spelling of the German word *fürchten* is displayed as [fʏːrçtən], and [[fʏrç|tən]] is the representation of this word in our implementation of the model.

2 Language and Evolution

In this chapter, we critically review the similarities and differences between biological evolution and language evolution. Our exposition concentrates on the most dominant effects that "change" species or languages and is, in part, rather simplistic. For a much more thorough discussion on biological evolution, see [11]. [15] contains many interesting articles highlighting specific aspects of the comparability of language evolution and biological evolution.

We first briefly describe the most fundamental principles of biological evolution. Organisms proliferate by creating offspring to which they transfer their genetic material. This material mostly consists of the DNA of the organism, i.e., long strings of nucleic acids arranged into chromosomes. The sum of all DNA of an organism is called its genome, which is contained equally in all cells of the organism. Evolution is a process which "operates" on the genome. It consists of two driving forces: Mutation and Selection.

[4] For further information on how to pronounce sounds in IPA–notation visit the website of Peter Ladefoged [21].

2.1 Mutation, Selection, and Adaptation

Mutations are random and erratic changes to the genome, which may result from nuclear radiation, errors in copying of genomes during cell division, or other factors. A functional mutation is one that impedes or enhances a certain biological function. Note that the only changes that count for evolution are those that are inherited to offspring. If an egg (or a sperm) cell is affected by a functional mutation, this mutation will be contained in all cells of the offspring of this particular cell. Thus, the offspring as a whole will be able to perform a certain function better or worse.

This change in function, a consequence of a change in a genome, is the target of selection. Selection in itself does not require any activity, but is a natural consequence of mutations affecting the ability of organisms to perform certain functions. For instance, a mutation that prevents an individual from extracting oxygen from the air is immediately lethal, i.e., it is prone to negative selection. Offspring carrying such a mutation will not live and will never reproduce, which lets the change itself disappear immediately. Similarly, any change that reduces the fertility of individuals will very likely die out within one or a few generations. Selection is therefore a two stage process: First, those individuals are positively (negatively) selected that perform an important function better (worse). Second, individuals which are themselves positively (negatively) selected have a higher (lower) chance of reproducing themselves, which lets the mutation spread (disappear) in a population.

However, most changes are not globally good or globally bad. Instead, their importance depends on the environment an individual lives in. For instance, an individual that is able to better balance its temperature when exposed to heat has an advantage in a hot environment, but no advantage in a cold environment. Thus, offspring carrying such a mutation that live in a hot environment will have an advantage. Over hundreds of thousands of years, these offspring will spread more successful than other individuals lacking the change, as they live longer, are less affected by heat strokes, and will thus generate more offspring. This process is called adaptation; it means the selection of random changes under environment-depending conditions.

2.2 Speciation

In many cases, reproduction requires two individuals to mate. This mating is not possible between any two living beings, but requires the partners to be somehow compatible. A group of individuals which are compatible to each other in the sense that they can create fertile offspring, and which are incompatible to all other individuals, is called a species[5].

Species are created when adaptation results in so many or such fundamental changes that incompatibility (with respect to reproduction) occurs. It is generally assumed that the most important cause for speciation is adaptation together with isolation. Assume that some

[5]Note that the term species only makes sense when sexual reproduction is involved.

ancestor has offspring that migrate to different regions which are isolated from each other, e.g., different continents or islands. As these offspring adapt to their new environment (by selection of random changes in their genomes), changes in their genome accumulate. If those changes are so many or so fundamental that the descendants of one offspring who migrated into a region are not compatible any more to the descendants of the offspring who migrated into another region, then a new species has emerged. Note that speciation usually only happens when the two regions are isolated from each other; if this is not the case, migration of individuals between the regions will result in a constant exchange of the adapted genomes.

2.3 Biological Evolution and Language Evolution

In summary, biological evolution is driven by random changes in the genetic material that is carried over to offspring. These changes survive in the offspring if they give them a selective advantage (or at least no disadvantage). As soon as the level of change exceeds a certain threshold, sexual incompatibility may result which leads to speciation. How well does this model fit to the development of languages? On first sight, the fit looks very good. A language is passed from parents to children. Languages evolve by changes in their lexicon or structure. If sufficient changes accumulate in a group of people, communication with other groups might get difficult, and eventually new languages emerge. Reflections about this toppic that are similar to our following deliberations can e.g. be found in [5, 23, 31].

However, a closer look reveals that the similarities actually are very limited, as that most phenomena in biological evolution cannot sensibly be carried over to languages. We believe that in the heart of the differences we describe now is the fact that in biology there is a clear distinction between the genotype and the phenotype of an individual. The genotype is passed to offspring and is subject to evolutionary changes which result in a phenotype being more or less suitable for selection. This fundamental difference cannot be drawn for languages. It manifests itself in a number of more fine-grained differences:

- *Genetic material:* In biology, the genome is the carrier of information between generations, and it is also the only target of evolutionary changes. In contrast, a language is not "inherited" from parents to children (though the ability to learn a language presumably is), but is instead learned by children through examples and education. This process does not only involve the parents, but also other persons a child is in contact with. Thus, it is not at all clear what should be the "target" to which random changes happen in languages.

- *Random changes:* Biological evolution happens in a random fashion and without direction; adaptation only appears through the combination of random change and selection. However, there is no sensible theory that would postulate that language change happens randomly. Instead, it is generally assumed that a change in a language, such as the way a particular word is pronounced, directly depends on some sort of advantage this change

gives to a speaker, such as a smoother way of speaking. Furthermore, it is completely unclear how a particular change that manifests itself is selected from the many ways to smoothen a language, and how the point in time when it manifests is determined.

- *Selection:* Random changes happen anytime and anywhere in genomes, but only those changes that increase the ability of an individual to reproduce itself survive; changes that are beneficial to an individual but that can not be inherited genetically are prone to disappear when the individual dies. There is no obvious way how this principle could be carried over to language change. Certainly, a language change does not survive by offering a speaker a greater chance to reproduce if he adopts it. Instead, language change is a social phenomenon. Changes that survive usually proliferate extremely fast and are adopted by all speakers of a language within very few generations. They are not only passed on to offspring, but also to other individuals within the same generation.

- *Adaptation and speciation:* The accumulation of sufficient changes within a genome may eventually lead to the creation of a new species. Within the realm of sexual reproduction, species boundaries are clearly defined by the ability to mate and produce offspring. The situation is quite different for languages. One may speak of English and German as two "language species" which are incompatible in the sense that a German (without knowledge of foreign languages) does not understand an Englishman and vice versa[6]. However, these two languages are only extreme points in a continuum. Other languages (or dialects), such as Plattdeutsch and Dutch, are placed in between. Historical linguistics tells as that for a long period of time, this continuum was so dense that one could have found a sequence of speakers of which each would have understood his neighbor, though the start person and the end person of the sequence would not have understood each other. To our knowledge, such a phenomenon is unknown in biology. Thus, species boundaries are not well defined for languages, and neither are speciation events.

- *The role of the environment:* Biological adaptation happens through the selection of certain changes that are beneficial in a given environment. This concept is not known in languages. There is nothing within a certain range on earth which would make one language more suitable for this range than another; languages do not adopt to temperature, food, height, etc.

- *Tree structure of speciation:* Genomic material is only passed from parents to offspring during reproduction. This process is generally called vertical transfer. In biology, there are only few cases where genetic material is not transferred vertically, with horizontal

[6]Note that we mostly compare languages with species, not with individuals. Both views have there merits; a detailed discussion of there respective advantages and disadvantages is beyond the scope of this report.

gene transfer in bacteria and hybridization of plants being the most notable ones. Vertical transfer naturally leads to a tree-model of evolution, where inner nodes of the tree represent speciation events. In contrast, the transfer of words between languages, which amounts to a horizontal transfer of information, is a very common phenomenon. Thus, tree models of languages are always somewhat inaccurate, as they ignore an important aspect of how languages evolve, i.e., by borrowing.

There exist a number of further differences which we mention only briefly. First, the time scale of evolution is measured in thousands of millions of years, while languages evolve much faster. Second, language change is strongly hindered by standardization processes on various levels, such as schools and school books defining what is wrong and what is correct, the standardization of the spelling of written language, national tendencies to increase the differences to other nations or populations, traditions manifesting in fixed sayings, songs, or prayers, etc. No such phenomena exist in biology. Third, language is a complex, multi-layered system with phonology, morphology, syntax, semantics and pragmatics. In contrast, DNA, although it has a certain degree of internal structure, is primarily a simple sequence of molecules which all have the same chances to be the target of evolutionary events.

All these differences may have consequences on how we should think about the reconstruction of language changes, which is the purpose of this paper. Languages seem to be much more versatile than species. The forces acting upon them seem to be much more random or erratic, because the highly conservative power of selection in biological evolution has no counterpart in languages. Furthermore, languages have no genes consisting of long strings of DNA. One could see words as the smallest unit of conservation, but words are comparably small (a few letters versus hundreds of nucleic acids) which strictly restricts their statistical potential. While two genes that are 90% identical form a strong phylogenetic signal, even words that differ by no means (homonyms) do not hint strongly to an evolutionary relationship. Therefore, word trees often are very far from true language trees, while gene trees often quite well approximate species trees.

Despite all these differences, we think there are several sensible arguments for studying the potential of phylogenetic algorithms from bioinformatics for their applicability to language. Many of the differences may only result in a small degree of "noise" while the dominant effects of language changes and how they accumulate prevail and can be recognized by algorithms. One also should consider that the tree model is not completely wrong even for languages. Ignoring the rare effect of creolization, languages do emerge by accumulating deviations from an ancestor language, and thus the backbone of language relationships still is a tree – though heavily overlaid with none-vertical relationships, and probably with a much less clear separation between children. A clear indication for the correctness of these thoughts is the fact that phylogenetic algorithms often reconstruct language relationships surprisingly well, at least regarding the topology of the relationships between languages.

3 Language change

Languages evolve in a variety of different manners. Especially the changes in the lexicon are well studied. New words arise to describe new technological or sociological attainments, words go extinct if they are not needed or used anymore. Words can also change their meanings or their pronunciation or they can be loaned from other languages. Changes also occur at other linguistic levels of a language and are not restricted only to its vocabulary, but may also involve syntax, morphology or even pragmatics.

In this work we focused on two different aspects of language change: sound change and borrowing. *Sound change* is the most common effect leading to a treelike phylogenie whereas borrowing requires a network model. The following brief introduction to the basics of phonetics and phonology follows [4].

3.1 Phonetic basics

The fundamental units of human languages are *phones*. A phone is the smallest sound segment and is a concrete realization of a *phoneme* [4]. Phonemes are the smallest units that distinguish meaning in a language (but itself don't have any meaning). For example /d/ and /n/ are German phonemes because they distinguish the difference in *Tod* and *Ton*. An average European language has about 30-50 phonemes [22, 24]. Each language has its own distinct inventory of phones and phonemes.

Phones can be characterized by articulatory characteristics, which is the basis of the *international phonetic alphabet* (IPA) [18]. The classification of the IPA distinguishes, among other things, pulmonic (that is egressive) consonants and vowels. Together, these two classes build almost all sounds of contemporary European languages.

A pulmonic consonant is characterized by a closure or stricture of the vocal tract sufficient to cause audible turbulence while letting out the air of the lungs. They can be distinguished by the following three properties:

- **Manner of articulation** (plosive, nasal, trill, tap/flap, fricative, lateral fricative, approximant, lateral approximant)

  Place of articulation (bilabial, labiodental, dental, alveolar, postalveolar, retroflex, palatal, velar, uvular, pharyngeal, glottal)

- **Voice** (voiced, unvoiced)

The pronunciation of vowels is characterized by an open pasage of the stream of air. The tongue changes its position but not its form to build different vowels. Vowels can be distinguished by the following four properties:

- **Vowel Height** (close, near–close, close–mid, mid, open–mid, near–open, open)

  Vowel Backness (front, near–front, central, near–back, back)

- **Vowel Rounding** (unrounded, rounded)

- **Vowel Length** (long, short)

Based on these properties, each vowel and each consonant can be represented by a vector that contains the particular aspects of the pronunciation of that phone. Vectors describing consonants have three, those describing vowels have four dimensions. For instance [d] is a plosive, alveolar, voiced consonant (see Figure A.1 in the appendix) and thus can be described by $(0, 3, 0)$, whereas $(6, 0, 0, 0)$ is the vector for the vowel [a]. The positions in these vectors correspond to the above properties in the mentioned order, the values refer to the feature lists of each property starting from zero.

3.2 Sound change

Sound change refers to processes that affect the phonological level of a language. Typical changes are *assimilation* and *dissimilation* [7].

Sound change often occurs in a way that one property in the vector describing a sound changes its value towards a neighboring value. If the new sound is not part of the set of phonemes of the language, one usualy assumes that it is skipped until an existing sound is reached. For consonants changes that affect the *voice* are far more common than those affecting, e.g., the manner or the place of articulation. Sound changes can occur *regularly* or *sporadic*. Regular sound changes apply to all sounds in the vocabulary of the language where the sound is within the same environment. On the other hand sporadic changes apply to only one or a few occurrences of the sound.

Assimilation refers to changes where the value of a property of a sound converges to the value of the same property in a neighboring sound in the word. This effect can be total (where the two values adjust to each other) or only partial. It can be related to a preceding or a following sound that is in direct or in distant contact to the changing sound. Assimilation is the most common effect in sound changes and is mostly explained with speech economy. *Dissimilation* is far more rare and refers to the opposite phenomenon, where two property values diverge from each other. It is often sporadic whereas assimilation is assumed to be usually regular.

3.3 Language contact

If two languages are in contact (or, more precisely, if speakers of these two languages are in contact) they may exchange linguistic characteristics. This exchange may occur on all linguistic levels. Most often words are borrowed from a foreign language. However, also phonemes that do not belong to the phoneme inventory of a language so far (usually in

[7] Other important aspects of sound change include monophthongization and diphthongization, epenthesis, metathesis, haplology and elisions and a lot more, all of which have not been implemented in our language change model so far.

combination with loanwords) or grammatical structures can be exchanged. A single word is commonly borrowed for one of two reasons: *demand* or *prestige*.

Demand Demand loanwords are words that are borrowed because the uptaking language lacks a proper word for a certain innovation (technological or sociological) in the borrowing language. *Yoghurt* (original turkish), *kayak* (inuit) or *algebra* (arabic) are loanwords in many languages.

Prestige Prestige loanwords are words that speakers of a language adopt because of the higher prestige of the foreign language. For instance French was a very prestigious language for centuries and many words have been borrowed from it (e.g. the word prestige itself in other European languages).

4 A formal model of language development

Based on the facts displayed in the preceding chapter we develope a formal model of lexical evolution of languages. The purpose of this model is not a complete implementation of all known effects of language change but a simple yet realistic working set that allows both treelike and networklike evolution.

This model is primarily used to generate artificial language families. During the simulation process all evolutionary events are logged. The log is used to generate the reference graph representing the phylogeny of the simulated languages. The vocabulary that has developed in the different languages is later passed to phylogenetic reconstruction algorithms. Finally, their outcomes are compared to the reference phylogeny to measure their quality under the given evolutionary parameters.

4.1 Basics

We place all processes into a landscape with a set of regions and neighborship relations between them. A landscape resembles a geographic region where groups of people can settle and migrate. A (part of a) population may migrate to a free neighboring region, taking their language with them. Their language may change by sound changes, two neighboring populations may exchange words of their languages or a population, and hence its language, can go extinct.

Geography All changes take place in a landscape of regions. Each of these regions has one or more neighbors. A region is called *active* if there is a population in it with its own language. Regions that are not active can be *populated* by neighboring active regions. This is modeled by copying the language of the region into the new one. So regions can be interpreted as placeholders that fill themselves with languages during the evolutionary process. At the beginning, there is only one active region populated by the root language of the simulation.

This geography is modeled as a graph. regions are nodes and neighborship between regions is represented by a directed edge. Every edge is linked with a `permeability` value that determines the probability for loaning a word along this edge (see Chapter 4.2.3). This value regulates the degree of contact between two neighboring regions to model low intensities of contact, e.g., due to a low prestige of a language or due to geographic conditions like seas or mountain ranges separating regions.

Vocabulary The root vocabulary consists of words for the 200 swadesh concepts. The *swadesh 200 list* contains 200 concepts, collected in the 1950ies by the american linguist Morris Swadesh, that are assumed to be essential to all human languages [36]. This means that every language around the world has to have a word for that concept (which may have been borrowed). Examples include body parts (like EYE, ARM or LEG), simple activities (like TO WALK, TO SEE or TO EAT), colors (like RED, GREEN or YELLOW), celestial bodies (SUN, MOON and STARS) or the numbers from one to five.

The count of 200 words is kept constant during our simulation as loanwords replace the existing word for the same concept.

Sounds As shown in Chapter 3.1, each sound can be described using a three or four dimensional vector. From the IPA 103 different sounds have been taken into account in this work, 45 vowels and 58 consonants. For technical reasons we sometimes differ from the IPA notation. Short vowels are displayed by minuscules and long vowels are displayed as majuscules (not with an attached [ː]). The english word of the swadesh concept *green* with the IPA–Notation [griːn] is thus represented as [[grIn]].

4.2 Evolutionary process

On the basis of the facts above (geography, sounds and vocabulary) the evolution of languages is simulated. The model uses an iterative approach. In every iteration one or more of the following events can occur independently for every active region:

- **Migration**: an active region migrates to an inactive neighboring region.

  Sound change: a sound change occurs.

- **Word transfer**: a word is loaned towards a neighboring region.

- **Extinction**: a language goes extinct (a region becomes inactive).

In the following we will discuss each of these events in more detail. A graphical representation of the process can be found in Figure 4.1.

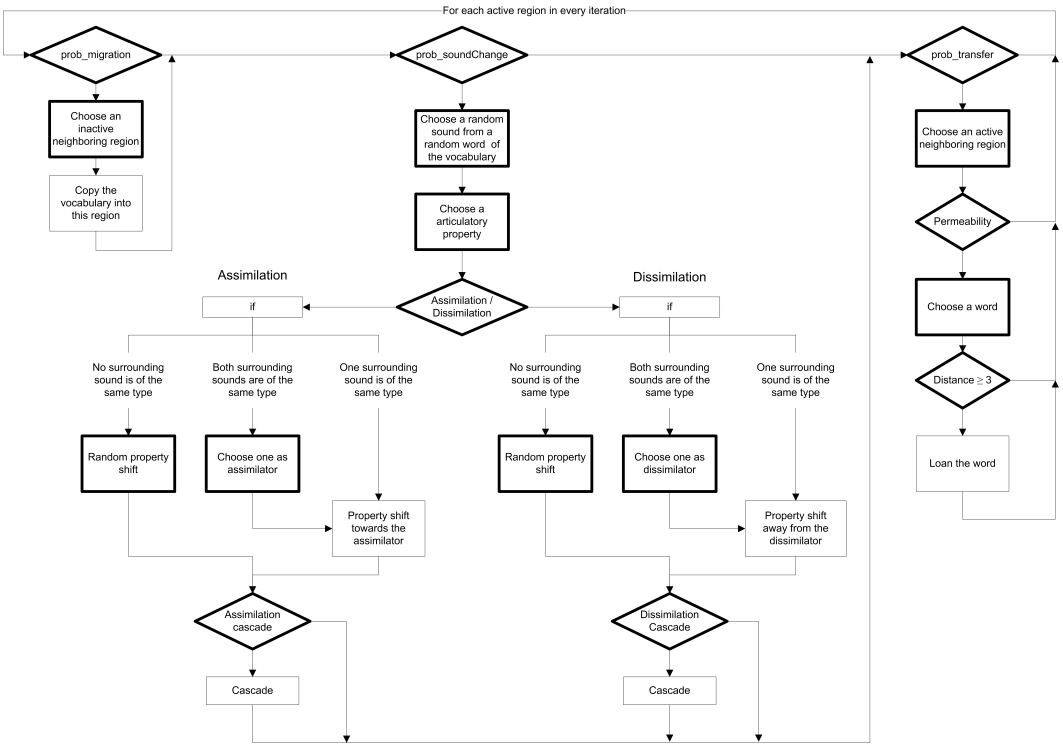

Figure 4.1: Schema of the processes in one iteration.

4.2.1 Migration

If a *migration* event occurs, an active region migrates to an inactive neighboring region by copying its current vocabulary to the new region. If no empty neighbor exists, nothing happens. Afterwards the new region is marked as active and an independent development for this region begins. The probability for this event to occur is controlled by a parameter (see Table 7.1).

Migration determines which region in which iteration migrates to which inactive neighboring region and thus determines the treelike backbone topology of the language relationships. This implies that if all regions are active no further migration is possible. Please note that there are two topologies to distinguish. The geographic topology of the regions and the topology of the phylogeny of the regions/languages. The fundamental influence of the first on the latter will be discussed later.

4.2.2 Sound change

A *sound change* event consists of two steps. First, a randomly chosen single sound of a single word in the regions vocabulary shifts towards a neighboring sound in the set of phones of a region. The probability for this event to occur is set in the `prob_soundchange` parameter. Afterward there is a chance that this sound change will be applied to all sound of the language in the same phonetic environment as the original sound.

Out of all sounds of the current region one is chosen randomly. Then one articulatory feature of this sound is chosen in which the change will occur. The `prob_soundchange_assimilation` parameter determines whether this change will be an assimilation or a dissimilation.

Assimilation / Dissimilation As described in Chapter 3.2 an *assimilation* is a sound change after which a sound in a word has converged towards another sound of the word, being a syntagmatic predecessor or successor of the sound. The original sound shifts the value of the chosen property towards the value of the other sound. If the resulting sound is not part of the language in that region the value is shifted until it results in a known sound.

A *dissimilation* works the same way like an assimilation, the only difference is that afterward the two sounds have diverged from each other.

Cascade (regular sound change) In a second step, after the actual sound change happened, the parameters `prob_soundChange_assimilation_cascade` and the corresponding `prob_soundChange_dissimilation_cascade` determine if this sound change is regular or sporadic. In case of a regular sound change all sounds in the same phonetic environment in any word of the language will change to the new sound too. Otherwise, the change is not adopted by other words.

4.2.3 Word transfer

For a *word–transfer* event, an active neighboring region of the current region along with a word are chosen randomly. If the region relationship passes the *permeability* threshold, which is set in the edge defining the neighborship of the two regions, the unweighted levenshtein distance of the loan and of the original word is calculated. If it is above a value of 2 the loan replaces the word in the other language. This threshold is used to distinguish a word transfer from an ordinary sound change.

The probability for the occurrence of this event is defined in the parameter `prob_transfer`. It influences the network edges of the phylogeny.

4.2.4 Extinction

If an *extinction* event occurs, a language goes extinct by marking its region as inactive and deleting the vocabulary of that language. Later active neighboring regions may (once again) migrate to this region. The probability for this event to occur is set in the `prob_extinction` parameter which among others effects the *starlike* value of a phylogeny which will be explained later.

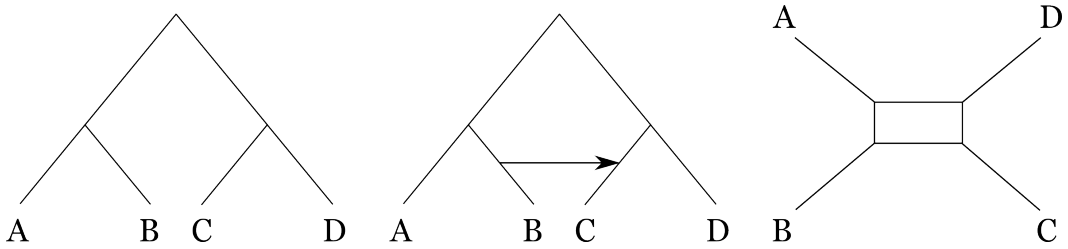

Figure 5.1: A phylogenetic tree, a reticulate network and a split network.

5 Phylogenetic algorithms

In the following, we give a brief introduction to the pyhogenetic algorithms we used in order to construct phylogentic trees and networks. To construct phylogenetic trees we chose *neighbor joining* and *maximum parsimony*, to construct split networks we chose *split decomposition*. Another interesting type of phylogenetic networks are reticulate networks (see Figure 5.1). Unfortunately we are not aware of any available algorithms/implementations that infer reticulate networks[8].

5.1 Tree algorithms

There exists a variety of different algorithms to reconstruct the phylogeny of a set of taxa based on a treelike model of evolution. They can be classified into two groups: Distance and character based methods. For a more detailed desciption of those methods, the reader is referred to [13].

5.1.1 Distance-based methods

Distance-based phylogenetic algorithms require a symmetric matrix D that contains the pairwise distances of all examined taxa. Gradually they construct a phylogenetic tree by melting two taxa and their corresponding lines and rows in the matrix. The methods differ in the way the two taxa are selected and in how their lines and rows in the matrix are joined. Since they are always joining two taxa they always create binary trees.

Neighbor joining is by far the most commonly used algorithm among the distance based methods and has proven itself very robust and fast in many studies. It was introduced in 1987 by Saitou and Nei [33]. Studier and Keppler reduced its runtime complexity to $O(n^3)$ in 1988 [35]. If the matrix is the exact representation of an additive binary tree the algorithm will exactly reconstruct this tree. The two pairwise taxa that are about to be melted are selected on a criterion that minimizes their distance and maximizes the distance of the two taxa to all other taxa in the matrix.

[8] We just recently took notice of [17] and will test the algorithm in the near future.

5.1.2 Character-based methods

Character-based phylogenetic methods do not operate on the reduced data of a distance matrix but on the characters that form the taxa themselves. Unlike tree methods they do not *create* a tree but *choose* the tree among all possible trees that explains the given data best under a given criterion. Each character is treated as a feature that evolves independently from all other features.

Maximum Parsimony is the most commonly used character based phylogenetic method. It searches for the tree of the given set of taxa that requires the minimal number of evolutionary events. The process of finding this tree/trees is divided into two problems, the small and the large parsimony problem.

Small parsimony deals with the problem of finding the labeling of the nodes of a tree, given a fixed topology, which explains the evolution of the taxa with a minimum of evolutionary changes/events. Labels for each inner node of the tree are generated that minimize the changes in the features towards the child nodes. Features that are equal in all taxa as well as features that differ in only one taxon are not informative and can be omitted.

The count of all necessary evolutionary events in the tree is called the *parsimony score*. Numerous methods have been presented to find the tree with minimal score. Among the most commons are the *Fitch* and the *Sankoff* algorithm as well as *weighted parsimony*. This problem has a runtime complexity of $O(nm)$ where n is the count of taxa and m the count of informative features, see [8].

Large parsimony deals with the problem of finding a topology *and* a labeling of the nodes that minimize the parsimony score. This problem is NP-hard, meaning that to find the optimal solution *every* possible solution has to be tested. Therefore, the parsimony score of every topology of an unrooted binary tree with n taxa has to be evaluated. As there are $\frac{(2n-5)!}{2^{n-2}(n-2)!}$ different unrooted binary tree topologies, the problem can not be solved exactly even for small taxa sets. Several heuristics have been developed that find a good (not necessary the optimal) solution in a sustainable time. Examples are the *branch and bound* and *nearest neighbor interchange* heuristics, see [8] and [27].

5.2 Network algorithms

Since many effects of language change can not be modeled in a phylogentic tree the construction of phylogenetic networks is an urgent topic. Several methods to construct phylogenies that do not evolve strictly along a tree have been suggested. Most of them are still in development and only very few are available as implemented algorithms. An overview of the different approaches can be found in [29] and [16]. The two most popular networks are *reticulate* networks and *split* networks.

Split networks compact representation of different incompatible trees that may explain the development of the taxa. A single node in a split network can not be considered as an

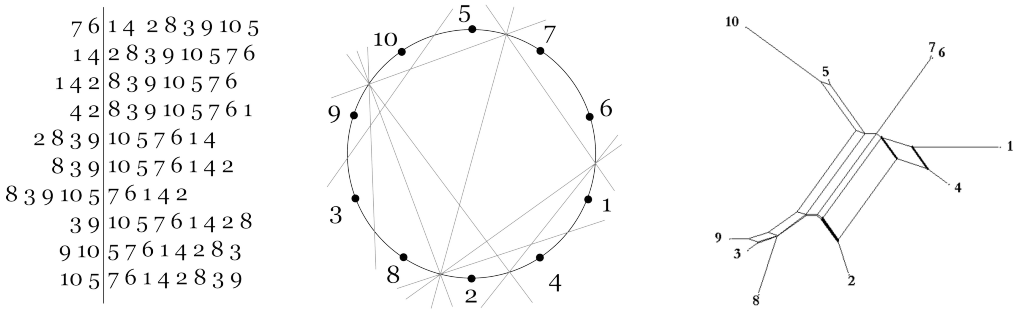

Figure 5.2: A *circular* set of 10 splits of a set of 10 taxa, their arrangement along a circle and a corresponding planar graph network (trivial splits excluded).

evolutionary event in the history of the taxa. Rather the structure of the graph identifies non treelike influences on the evolution.

Reticulate networks try to reconstruct graphs that represent the history of the taxa. By inserting directed edges to the tree, effects like borrowing or creolization (or their biologic counterparts of horizontal gene transfer and hybridization) are modeled. The underlying methods are more difficult than those for split networks and to the best of our knowledge, no implementation is available.

5.3 Split decomposition

Split decomposition is a distance based algorithm to construct split networks. Although a split network does not represent a history of taxa it is constructed in a way that the sum of the edges along the shortest path between two labeled nodes is proportional to the distance of their corresponding taxa in the distance matrix. Therefore, often a certain extend of genealogic structures is recognizable.

The algorithm that calculates a split decomposition network operates in two phases. First it uses the distance matrix to calculate bipartitions of the set of taxa (splits) and thereafter uses them to construct the network (see [16] for details).

Bandelt and Dress introduced the concept of *circular* split systems [1]. A set of splits is called circular if there exists an arrangement of the taxa in the set along a circle so that every split of the set can be represented by a secant through that circle (see figure 5.2). These sets do have a planar split graph representation [6]. Sets of splits that are calculated by the split decomposition algorithm are circular and hence yield planar graphs. If no secants cross each other inside of the circle there are no contradictions and hence the resulting graph will be a tree and, vice versa, every tree can be represented by a set of circular splits with no crossing secants.

5.4 Distances of phylogenetic graphs

To validate the quality of the phylogenetic algorithms under different settings, a measure for the similarity (or the distance) of phylogenetic graphs has to be defined. This measure is then used to compute the distance of the inferred graph to a reference graph, this graph represents the language relationships and is generated from the knowledge of the simulation process.

Since our model for language change incorporates language contact this graph would naturally be a network. But a reference split network can not be generated because nodes and edges in a split network do not represent concrete events. They may only be used to find clues for non treelike events, see chapter 5.3. Hence our reference is still a tree which is compared to both, trees and split networks.

5.4.1 Robinson Foulds tree distance

Bourque (1978; [3]) and later on Robinson and Foulds (1981; [32]) developed a distance measure for trees with the same count of leafs. It is based on the count of edges in which the two trees differ.

Unrooted trees can be represented as sets of bipartitions, called splits. Every edge in a tree divides the leafs of the tree into two disjoint sets that are in one connected component after the removal of this edge. A split is called trivial if it separates only one taxon from the others. The *robinson foulds* (RF) distance of the two trees is defined as the number of different non trivial splits between them.

This distance measure is easy to implement but has one major drawback. It is very sensitive to small differences between the trees and punishes differences more than it honors similarities. Figure 5.3 shows an example in which the two trees share a major part of their topology but do have the greatest possible RF distance to each other, because they do not have any exact split in common.

Upper boundary To compare the RF distances of different graphs that do not have the same number of taxa it is necessary to evaluate an upper boundary to which the two distances will be normalized. Under the condition that the trees are unrooted phylogenetic (binary) trees with n taxa it can be derived as followed:

> An induction easily shows that an unrooted binary tree always consists of $n-2$ inner nodes (each with 3 edges) and $2n-3$ edges. Since there are n leafs there are n trivial splits. This leaves $n-3$ non trivial splits in each of the two trees. If they do not have any single split in common every tree yields $n-3$ to the distance resulting in the maximum RF distance of $2n-6$.

There exists a variety of other tree distance measures like the *Kuhner Felsenshtein* distance, the *quartets* distance or the *tree edit* distance, each with its own drawbacks and shortcomings. They haven't been used in this work.

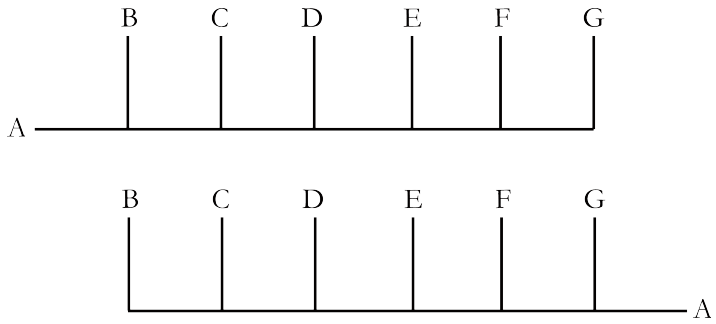

Figure 5.3: This two trees share a great deal of their topology but they do have the greatest possible RF distance.

5.4.2 Robinson Foulds split network distance

Since split networks actually are representations of multiple trees, the robinson foulds approach can be used to measure the distance between two split networks. Furthermore, since a tree is a special case of a split networks, it also can be used to measure the distance between a phylogenetic tree and a split network.

Trees as well as splits networks can be described by a set of bipartitions or splits. The count of splits that is equal in the two networks (or the tree) is the analogue distance to the RF distance of trees.

Upper boundary Under the condition that one of the examined graph is a tree and the other is a circular split network a upper boundary for their distance can be derived as followed:

> The tree has exactly $n-3$ non trivial splits. Using again an induction it can be shown that a circular set of split can only contain maximal $\frac{n}{2}(n-1)$ splits. n of these splits are trivial leaving $\frac{n}{2}(n-1)-n$ non trivial splits. if the two graphs have no split in common the upper boundary for their RF analogue distance is
>
> $$\left(\frac{n}{2}(n-1)-n\right)+(n-3) \quad = \quad \frac{n}{2}(n-1)-3 \tag{5.1}$$
>
> This theoretically possible boundary was never roughly reached in the experiments. This suggests that there exists a lower upper boundary for graphs of these types that we were unable to determine yet.

5.4.3 Robinson Foulds split network similarity

Similar to the RF distance an RF similarity can be defined as the number of equal splits of two graphs. The number of non trivial splits in two unrooted binary trees A and B is always constant. Together they have $2n-6$ non trivial splits assuming that they both have n leafs. The count of equal splits in both tree (which would be their similarity $s_{A,B}$) can be calculated by their RF distance $d_{A,B}$ as follows:

$$s_{A,B} = \frac{2n-6-d_{A,B}}{2} \tag{5.2}$$

Thus, concerning two trees, this measure does not contain any new information about the trees. The count of non trivial splits in split networks however is *not* constant. Therefore the count of identical splits in both graphs contains information about them that exceeds their distance. Under the condition that one of both graphs is a tree, $n-3$ is an upper boundary for this similarity. The similarity measure for a phylogenetic tree and a splits network would thus be the count of equal splits in both graphs normalized to the upper boundary of $n-3$. Note that in a mathematical sense this is not a strict similarity measure since a similarity of 1 does not mean that the two graphs are identical since they still can have a distance greater than 0.

6 Reconstruction process

After the simulation of a language evolution under the presented model, the created languages are analyzed by the phylogenetic algorithms. The inferred graphs are compared with the reference phylogeny constructed from knowledge of the simulation process. In the following, this processes is explained in more detail.

6.1 Substitution matrix of sounds

In this work the distance between two words, which is necessary to infer a phylogeny, is computed as their weighted levenshtein distance. This requires a substitution matrix of sounds, which should represent the evolutionary distances of sounds. The computation of such a matrix can be reduced to finding shortest paths in a graph. Therefore a graph with all sounds is created which represents the phonetic neighborship of sounds. Two sounds are phonetic neighbors if their properties vectors differ in just one value and the two different values are closest to each other in the given combination of the other properties.

Since the sound changes occur only along the edges of this graph, the greater the distance of two sounds in this graph, the greater their evolutionary distance. In this manner all pairwise distances of sounds were calculated and stored in the substitution matrix.

6.2 Data preparation

Before the phylogenetic reconstruction can begin some processing of the generated data is to be done to provide all the necessary basics for the algorithms.

Reference phylogeny After a simulation has finished a phylogenetic tree is generated using the recorded entries for the occurred treelike events *migration* and *sound change*.

Starting with the active regions at the begin of the evolution each *sound change* event extends the leaf edge to the affected region by 1. If a *migration* event has occurred, the edge is ended and an inner node and two new edges are created. Every subsequent sound change extends the edges of the two languages independently.

Distance matrix of languages To construct the distance matrix of the generated languages the weighted levenshtein distance is used. To measure the distance of two languages the pairwise distances of the 200 swadesh concepts in the vocabulary of the languages are calculated and summed up.

6.3 Phylogenetic reconstruction

Three algorithms are tested in the phylogenetic reconstruction process, the distance based method *neighbor joining*, the character based *maximum parsimony* and the (distance based) network algorithm *split decomposition*. We use the software package `SplitsTree` of D. Huson and D. Bryant in version 4.2 which implements all of these algorithms.

The character based *maximum parsimony* is not calculated by `SplitsTree` itself, but by the *dnapars* program of the *phylip* package of J. Felsenstein [9]. This program unfortunately does not allow the usage of any desired character but is limited to the A, G, C, T alphabet of DNA strands. Therefore the IPA letters had to been encoded into this alphabet.

6.4 Distance measure

The quality of the inferred graphs has to be validated. The distance between the reference phylogeny and the graph is measured using the method described in Section 5.4. The determined values are stored along with the parameters for this evolutionary run. The log is used the run comparative analyses in relation to varying evolutionary parameters.

7 Results

In this chapter, we report on the results of various experiments we conducted for judging the ability of the three phylogenetic algorithms explained above to reconstruct a simulated history of languages. We run simulations with a multitude of different parameter settings, logged the true events of language change and migration, and compared the RF-distance of the reconstructed graphs to the recorded reference graph. We concentrate on the three following questions:

- To which extend does borrowing of words deteriorate the performance of the different algorithms?
- In which way does the structure of the topology of regions influence their performance?
- What is the influence of the length of the branches, i.e., the amount of independent evolution in a language without speciation events?

The results we report are preliminary and require further study. In particular, they are constrained by the lack of an adequate algorithm for modeling and reconstructing evolutionary networks. Due to this lack, the reference graph itself is a tree. Although this tree will be

iterations	2000
prob_extinction	0 ‰
prob_migration	2 ‰
prob_soundChange	200 ‰
prob_soundChange_assimilation	90 %
prob_soundChange_assimilation_cascade	90 %
prob_soundChange_dissimilation_cascade	20 %
seed_evolution	NULL
seed_transfer	NULL
seed_topology	15335

Table 7.1: Fixed parameters in our experiments.

the true tree, describing how languages emerged from each other, it lacks the "network-like" information of the true graph of phylogenetic relationships. Thus, we actually study the ability of different algorithm to recover the backbone tree in a phylogenetic network, rather than their ability to recover the network itself. However, we believe that answering this question has merit in itself.

Note that the space of possible parameter settings in our model is enormous. We cannot hope to explore this space in an exhaustive manner. Therefore, in our experiments, we usually keep all parameters constant except one. The values we used are shown in Table 7.1. If not specified otherwise, we used a landscape of 10 intermediately connected regions. We also plan to explore combinations of parameter changes in future work.

To judge the statistical robustness of our results, we also computed the RF-distances between the outcomes of *neighbor joining* of a random distance matrix and the reference trees. In this runs, the normalized average distance was 0.98 with a standard deviation of just 0.02. Thus, the trees generated from random data in almost all cases had the greatest possible RF distance. In contrast, the reconstruction algorithms obtained much lower distances, meaning that they are able to detect a phylogenetic signal even in extremely noisy data (see below).

7.1 Language contact

As a first experiment, we examined the effect of increasing language contact on the performance of the phylogenetic methods under study. We varied the parameter `prob_transfer` between 0 and 1000 ‰. A value of 0 means that no contact at all took place, while 1000 means that every language tries to loan a word toward an other language in every iteration, which might not succeed (see section 4.2.3). We expected a deterioration of the distances of the reconstructed graphs to the reference graph, in particular for all tree-based methods, because a higher degree of language contact yields an increasing amount of signals that are not compatible with treelike evolution.

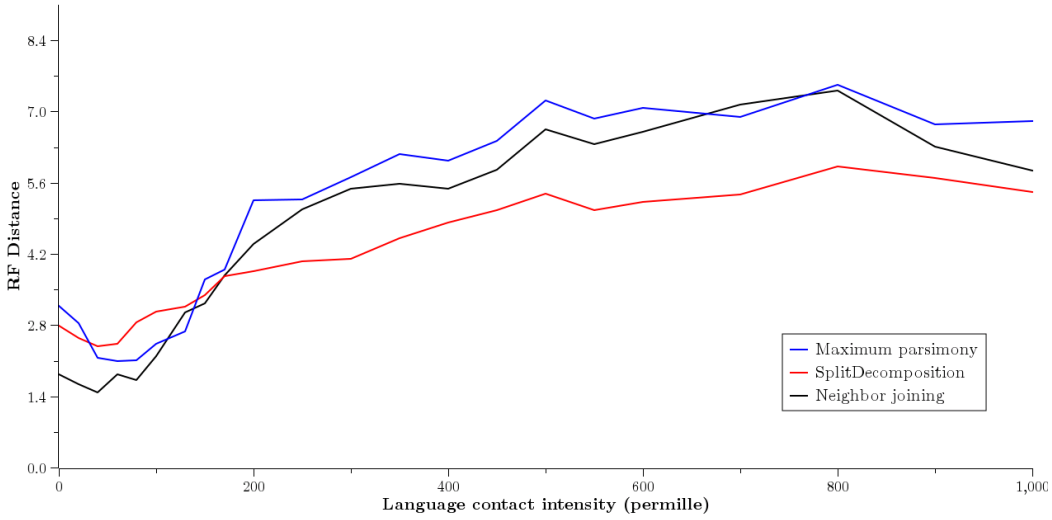

Figure 7.1: The RF distance of the inferred graphs to the reference graph in dependence of the language contact intensity.

Figure 7.1 shows the RF-distance of the reconstructed graphs to the reference graph for increasing values of `prob_transfer`. After a short decrease in the distance, meaning an increase in reconstruction quality, the expected increase in quality can be observed. This increase levels out at about 500‰. *SplitDecomposition* starts with a higher distance than the other two methods but the following increase is less steep. From a value about 150‰ on, it outperforms the other methods. We will discuss the decrease at the lower contact intensities in Chapter 8.

As stated in Chapter 5.4.3, the RF-similarity between the reference and the split graphs is not simply reciprocal to their RF-distance. Therefore, we measured the similarity between the different graphs in addition to their distance. The results are shown in Figure 7.2. The general progression is expectedly inverse to the distance. Interestingly, *SplitDecomposition* always has a lower similarity to the reference than the results of the tree-based methods – even when the observed RF-distance is lower. This implies that this algorithm creates graphs that agree *and* disagree in fewer splits to the reference than those of the other algorithms.

7.2 Geographic topology

In a second experiment, we analyzed the influence of the underlying geographic topology. Note that this topology must have a very strong impact on the resulting languages. Imagine a landscape where separate regions are connected in the form of a directed, rooted tree, and that the first active region is the root of the tree. In such a world, language evolution will be almost exclusively tree-like. There will still be other influences, because parent languages may loan word to child languages (but not vice-versa when the edges are unidirectional), and because languages may go extinct which opens room for new migrations. In the other extreme, imagine a landscape where all regions are reachable from all other regions with equal

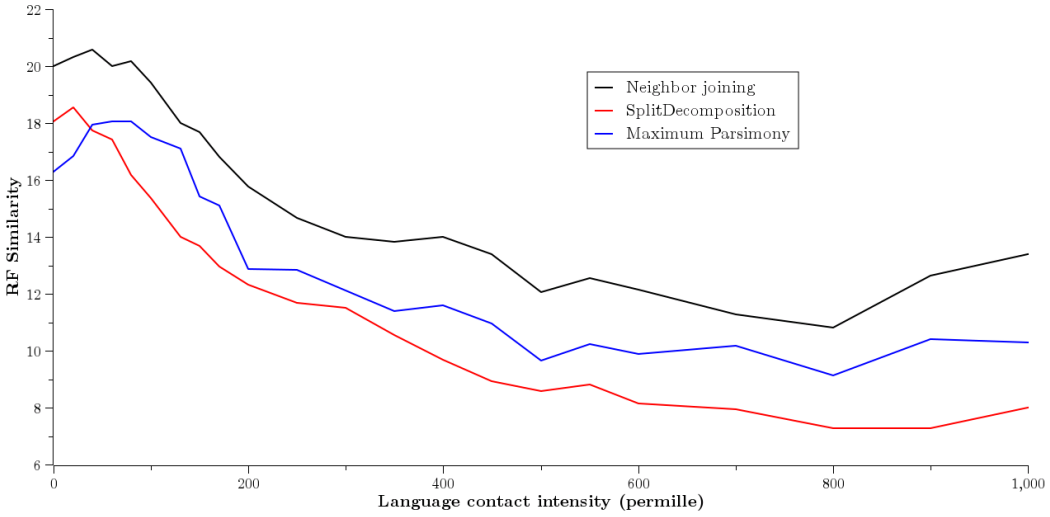

Figure 7.2: The normalized similarity of the inferred phylogenies to the reference graphs.

ease. In this model, there will be a point in time when all regions have been settled, and from this point on, the only changes that can happen are loaning words (ignoring extinction). Very likely, this borrowing will completely outweigh the initial tree-like signal if the simulation is run long enough[9].

To study the influence of the geographic topology on the inferred phylogenies we ran experiments with 10 regions arranged as depicted in Figure 7.3. We expected the following behavior:

- In the ring model, migration can only (ignoring extinction) happen in a linear fashion. Contact between languages can only happen along edges that have also been used for vertical evolution, hence the network effects should not distract the tree-algorithms to the same degree as in the other two models.
- In the fully-connected model, migration will be fast and will mimic an evolutionary tree, but the tree signal will later be heavily overlaid with contact between all languages. We expect that, depending on the length of the simulation, the tree-based algorithms should perform much worse than the network-based method.
- In the sparsely connected, planar topology, migration should be slower, and the effect of language contact has a more local flavor. Thus, we expect the performance difference between tree-based and network-based algorithms to be smaller than in the fully connected model.

Figure 7.4 shows the RF distances of neighbor joining of the three geographic models, Figure 7.4 does the same for SplitDecomposition. In general, our expectations regarding

[9] Note that a similar argument – the increasing mobility of a large fraction of worlds inhabitants – is frequently used today to explain the nowadays observed rapid decrease in the number of spoken languages in the world.

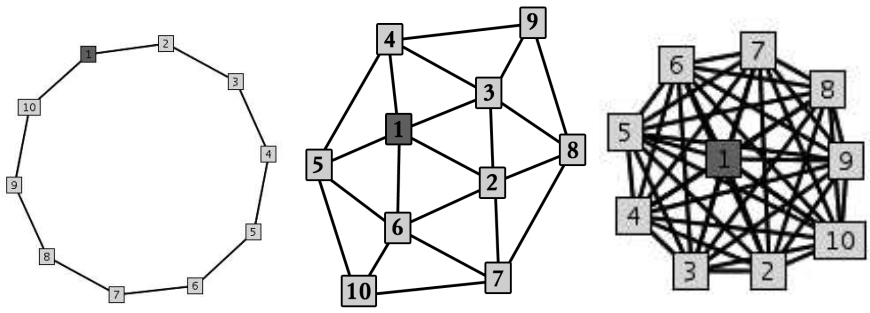

Figure 7.3: The three studied geographic topologies with 10 regions.

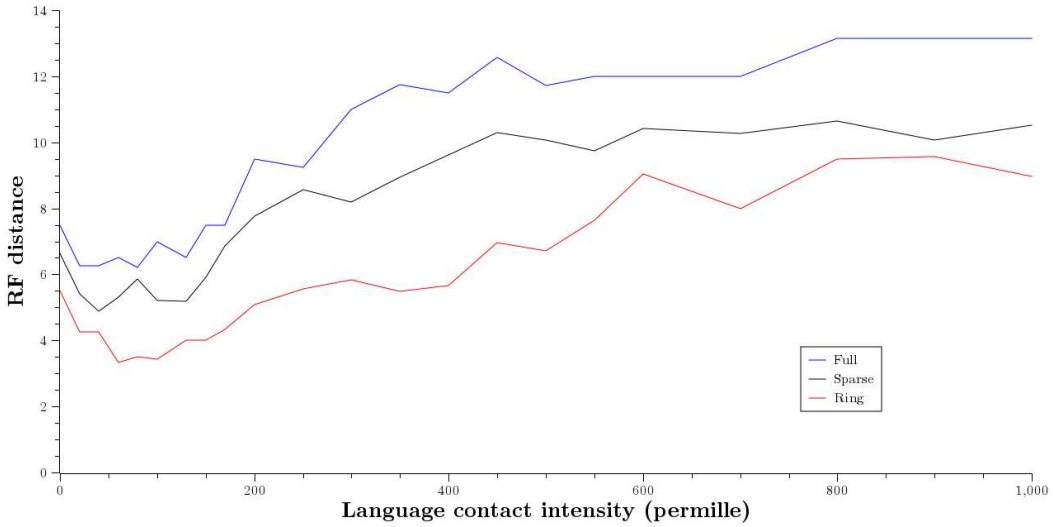

Figure 7.4: The RF distances of neighbor joining in the three geographic models.

the performance of the tree-algorithm neighbor joining have been confirmed. In the circular circular geography the algorithm performs much better than in the the sparsely connected or fully connected environment.

The increase in inference quality at lower contact intensities occurs in all environments. Yet the degree and duration of increase differ. When there are less connections between the regions (ring) the increase lasts longer and is stronger as in the fully connected geography.

SplitDecomposition shows a quite erratic behavior. In the circular environment it performs very bad and the RF distance almost reaches its upper boundary. the sparsely connected geography yields almost constant RF distances but at a higher level than in the before mentioned experiments (probably due to a higher *starlike* value, see Section 7.3). In the fully connected geography the RF distance decreases with rising contact intensity, which is completely counter-intuitive.

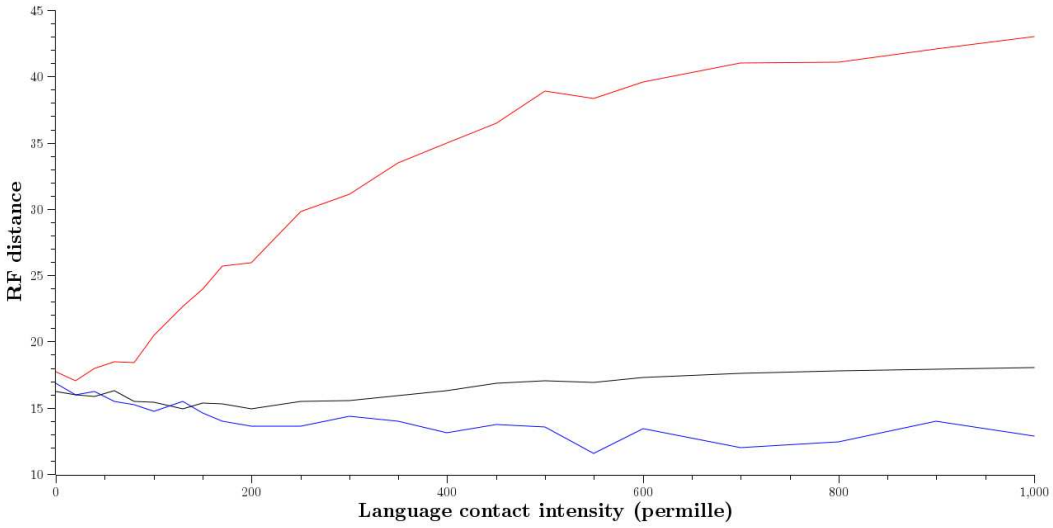

Figure 7.5: The RF distances of SplitDecomposition in the three geographic models.

7.3 Branch lengths of the reference tree

Finally, we examined the influence of a particular property of the reference tree on the quality of the reconstructed graphs. Since our geographic model is finite, the effects of language change become more and more dominant compared to the genealogic relationships the longer a simulation runs. In the reference tree, this is reflected in the proportion of the lengths of the edges between inner nodes and the lengths of the edges from inner nodes to leaves. We call this proportion the *starlikeness* of a graph and compute it using the following formula:

$$\frac{\sum \text{length of leaf branches}}{\sum \text{length of all branches}}$$

Clearly, this measure is primarily influenced by the number of `iterations`, the probability of migration events `prob_migration` and the rate by which languages go extinct, `prob_extinction`. The influence of `iterations` is obvious since more time in an evolutionary run results in longer branches. `prob_migration` indirectly determines at what time all regions are occupied (meaning the end of migration events), and `prob_extinction` potentially empties an active region leading to a new migration in late stages of the evolution.

Figure 7.6 shows the RF distances of the neighbor joining trees versus the contact intensities at different starlike measure values. One can observe that for topologies with a starlike factor of about 60% (this was the lowest value in our simulations), this algorithm constructs trees that are very close to the reference up to a contact intensity of about 200‰. An increasing starlike factor heavily decreases the performance of the algorithm. From a value of 90% starlikeness on, the reconstructed tree is not distinguishable any more from a pure random tree.

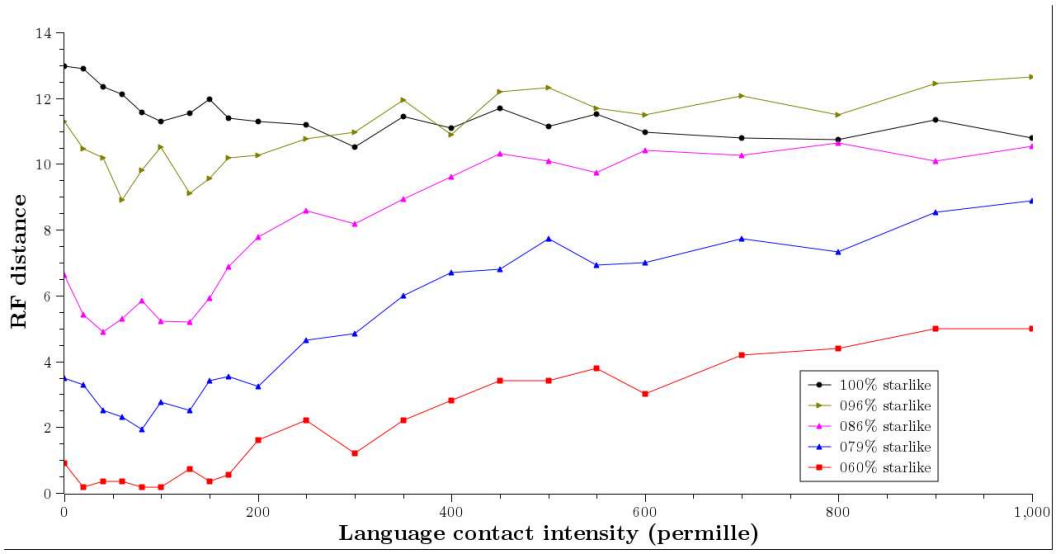

Figure 7.6: The RF distance of the neighbor joining algorithm to the reference phylogeny in dependence of the contact intensity and the *starlike* factor of the reference tree.

Figure 7.7 shows the RF distances of the SplitNetworks to the reference tree versus the contact intensities at the same starlike values as the neighbor joining figure. This algorithm shows no clear dependence from the starlike factor of the reference graph. Although the RF distances differ markably there is no obvious correlation between them and the starlikeness.

8 Discussion

The experiments reported in Chapter 7 clearly show that the quality of the inferred graphs of all examined algorithms decreases with increasing language contact intensity. Starting from a simulation without language contact, there is an – at first sight surprising, but see below – temporary increase in quality of reconstruction in the contact range from 0 to about 60‰. This is followed by a continuous decrease in quality until at about 500‰, from which on the quality remains roughly the same. This is due to the saturation of the average share of loanwords in the simulated languages.

At reasonable starlike factors of the reference graph, the inferred topologies always are considerably closer to the reference than graphs that have been calculated using a randomly chosen distance matrix. Even in extreme cases, where languages end with an average share of loanwords of up to 90%, the examined algorithms are still able to infer correctly parts of the originate phylogeny. This may be explained as follows: If a word is borrowed in an early stage of the evolutionary process, the following evolutionary events will happen to this word as to any other word in the language, and thus make this word become closer to the other words of the language – independently of the development of the same word in its original language. Therefore, such loanword in the end also carry information about genealogical relationships.

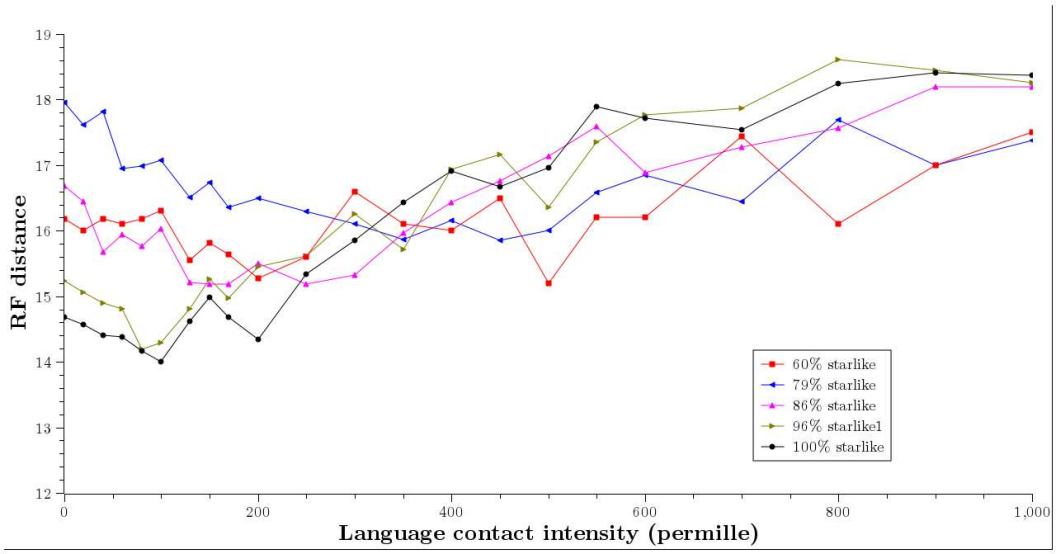

Figure 7.7: The RF distance of the SplitDecomposition algorithm to the reference phylogeny in dependence of the contact intensity and the *starlike* factor of the reference tree.

Maximum parsimony Conceptually, *maximum parsimony* (MP) should be superior to the *neighbor joining* (NJ) method because it uses more information. However, in our experiments MP consistently infers trees that are a bit less close to the reference than those creates by NJ. By close inspection, we found that this effect actually might be an artifact of our currently used implementation. Recall that we had to encode all sounds in the DNA alphabet to be able to use the MP implementation in the SplitsTree package. However, our current mapping of sounds to DNA contradicts the assumptions of the MP method. In our implementation, up to four evolutionary events (instead of one) may be necessary to change a sound. For instance, the two distant sounds [ʃ] and [ɳ] received the very similar codings of *ACCA* and *ATCA*, respectively, with a edit distance of just 1, whereas the sound [ʒ] is encoded as *GAAT* and thus four evolutionary events away from the very similar [ʃ]. Actually, we find it rather astounding that even under this conditions the algorithm still competes quite well.

There are two ways to solve this problem. First, the coding of the sounds in the DNA alphabet can be optimized such that similar sounds get similar codings and distant sounds get distant codings. A simpler solution would be to shift to a different MP implementation that is able to work on any set of characters.

SplitDecomposition The SplitDecomposition (SD) algorithm showed a very erratic behavior in our experiments. In our first experiment it seemed promising at higher contact intensities. Although at low contact intensities the SplitNetworks were more distant from the reference graph than the inferred trees of the other algorithms, this changed at about 150‰ (which amounts to an average share of loans of about 60%).

However this does not hold true in the last two series of experiments. Here the algorithm shows a different and somewhat unpredictable behavior. While NJ and MP always show the

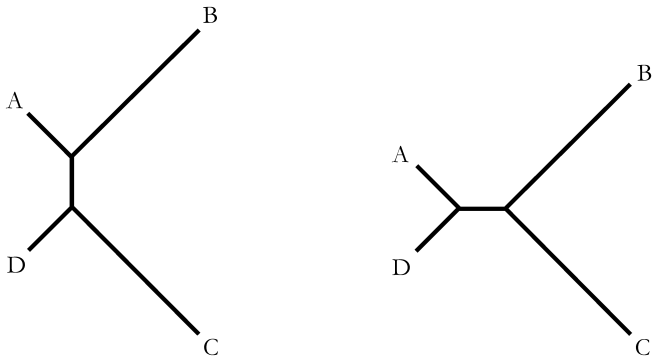

Figure 8.1: Long branch attraction: Phylogenetic algorithms commonly arrange phylogenies like in (a) incorrectly by grouping the long branches together like in (b) due to spurious similarity that appears when sufficient independent change happened.

same general progression in their RF distances (a first decrease in the distances, followed by an increase and a saturation) SD either rises constantly, stays almost the same, shows the same progression as the other two methods or might even be decreasing. This problem, along with its structural dilemma of the very hard to interpret result graphs, leads us to the conclusion that this method is not very well suited for our purpose, i.e., the inference of historic language relationships.

Geographic topology The geographic topology massively influences the outcome of the reconstruction algorithms. If there are just few connections between the regions, the overall reconstruction quality is better than in stronger connected geographies. In case of a circular geography the results are best whereas a fully connected geography yields the worst results. This can be explained by the fact that in loosely connected geographies the paths along which language contact happens is more likely to be the same along which the migration happened. In a circular geography, language contact can appear between parent-child-languages only, which does not interfere with the treelike evolution. In fully connected geographies, on the other hand, every language can exchange words with any other language, closely related or very far related languages alike.

Reference tree The analysis of the influence of the *starlike* measure on the quality of the inferred graphs shows that an increasing starlikeness leads to significantly worse reconstruction results. The lowest achieved starlike value of about 65% yields the best reconstruction results whereas starting from about 90% the infered graphs reach the distance level of random graphs.

This is due to fact, that the share of phylogenetic informative data is more and more reduced to the inner part of the tree. The longer the leaf branches are the more distracting or useless information appears in the vocabulary of the languages. Every sound change event that occurs since the last migration event of a taxa does not contain information about the

(treelike) phylogeny of the set of taxa. The sound change occurs in this particular branch only, or it is borrowed towards another language and therefore interferes with the data containing information about the treelike phylogeny.

Long–branch hypothesis The tree algorithms show an increase in quality (or a decrease in the distance) at small contact intensities. This implies that the algorithms do not infer the best results when there is no language contact but when there is just a little. This unexpected behavior may be explained as follows. The algorithms are well known to be subject to an effect called *long branch attraction* (LBA) [8, 7, 14]. This term denotes the (false) behavior of phylogenetic algorithms to group taxa with long branches together, which contradicts the underlying phylogeny. Figure 8.1 demonstrates this behavior on a small example phylogeny. SplitDecomposition supposedly does not suffer from this effect, or at least not in to the same degree as the other methods [2], does not necessarily show the explained progression.

We postulate that in our model the influence of LBA is the strongest when there is no contact at all (`prob_transfer` $= 0$) and decreases with rising contact intensities. This positive development is superposed by the rising negative influence of higher loan shares in the vocabulary of the languages.

To provide evidence for this assumption, we measured the average branch lengths in our reference trees as well as their variance (see Figure 8.2 for the variance). Both decrease with rising contact intensity, which is a side-effect of the rising share of loan words since a higher loan share leads to more similar vocabularies in neighboring regions which reduces the distance of the regions. This means that rising contact intensity yields smaller distances between languages which leads to shorter branch lengths as well as smaller variances. This reduces the influence of LBA which positively affects the infer quality. However, from a certain degree of contact on the negative influence of the contradicting signal surmounts the improvement of the decreasing influence of LBA, which together results in the expected deterioration of reconstruction quality.

One can actually fit a numeric model of the two influences to mimic the change in quality. Let $avgB$ be the average share of loans and let $\sigma(\mathrm{BL})$ be the variance of branch lengths. Then, the following function is very close to the observed behavior (see Figure 8.2):

$$0.2\ \sigma(\mathrm{BL})^4 + 0.8\ \mathrm{avgB}^4 \tag{8.1}$$

Note that the formula is only an illustration, and that the coefficients in this formula originate from $a-posteriori$ correlation and not from analytical considerations. The concrete values also depend heavily on various properties of our models, such as the starlike factor.

9 Summary and Outlook

In this work we investigated the behavior of phylogenetic algorithms under a variety of different settings in a simulation environment. We presented a formal model of language change that

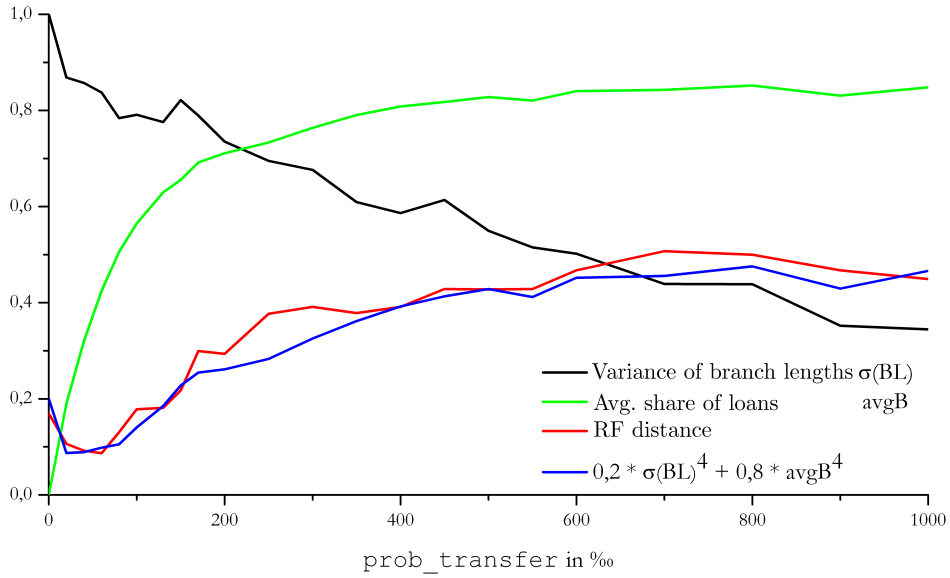

Figure 8.2: The approximation of a curve to the behavior of the RF distance by superposition of the average share of loans (avgB) and the variance of the branch lengths σ(BL).

includes tree as well as network based events. We used this model to investigate the ability of different algorithms to infer the simulated phylogeny under different language contact intensities. Our results show that a rising level of language contact yields a decrease in reconstruction quality in all examined algorithms. The tree-based algorithms have proven very robust to network influences, whereas the examined network-based algorithm sometimes showed a unpredictable and almost essentially erratic behavior. However, for a certain range of parameters, it outperforms the tree-based methods. Our experiments also show that a small degree of contact actually improves the quality of all examined algorithms, which can be explained by the influence of the long branch attraction.

We also showed that when interpreting the results of phylogenetic algorithms several surrounding parameters of the phylogeny have to be taken into consideration carefully. This includes the geographic conditions under which the (language) changes have taken place as well as properties of the graphs themselves.

Open questions and future work This study also left a number of open questions for future research. First, our current implementation of the MP must be replaced to allow for a more fair comparison. Second, we observed that the field of phylogenetic network algorithms is still in its infancy, probably because the respective effects are rare in biology. However, as they are everywhere in language, we believe that this line of research must be intensified to provide more robust models for inferring language relationships. In particular, the SplitDecomposition method has proven to behave very unreliable, adding to the intrinsic problems of their interpretability. We plan to consider other types of network reconstruction algorithms in future work, such as the recently presented *galled networks* [17, 19, 20]. Adopting a more

suitable network model would also allow us to move to more suitable forms of measuring the distance between inferred phylogenies.

As a second line of future work, we plan to extend our model of language change to account more linguistic events, such as creolization. Also the linguistic levels on which the algorithms works will be extended beyond the lexicon. We plan to adopt our algorithms to linguistic data on all levels, as they are available in richly annotated corpora.

Finally, it is obvious that our model settings are mostly far from adequately describing a realistic environment. It would be very interesting to try to fit our model parameter to those obtained from a real language history. This should particularly affect the relative frequencies of evolutionary effects, the geographic topology, and the migration probabilities.

References

[1] H.J. Bandelt and A. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92:47–105, 1992.

[2] J. Bergsten. A review of long-branch attraction. *Cladistics*, 21(2):163–193, 2005.

[3] M. Bourque. *Arbres de Steiner et reseaux dont certains sommets sont e localisation variable.* PhD thesis, University de Montreal, Montreal, Quebec, 1978.

[4] J. Clark and C. Yallop. *An Introduction to Phonetics and Phonology.* Blackwell Oxford UK & Cambridge USA, 2. edition, 1995.

[5] W. Croft. *Explaining Language Change, An Evolutionary Approach.* Pearson Education, 1 edition, 2000.

[6] A. Dress and D. Huson. Computing phylogenetic networks from split systems. Manuscript, 1998.

[7] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, Vol. 27, No. 4, 1978.

[8] J. Felsenstein. *Inferring Phylogenies.* Sinauer Associates, Inc., 2004.

[9] J. Felsenstein. Phylip v3.65. http://evolution.gs.washington.edu/phylip.html, 2005-08-30.

[10] W.T. Fitch. Linguistics: An invisible hand. *Nature*, 449(7163):665–667, 2007.

[11] S.J. Gould. *The Structure of Evolutionary Theory.* Belknap Press, 2002.

[12] R. D. Gray and Q. D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965), 2003.

[13] D. Gusfield. *Algorithms on strings, trees, and sequences, Computer science and computational biology.* Cambrigde University Press, 2 edition, 1999.

[14] Z. Hang. Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution*, 42, 1996.

[15] G. Hauska. *Gene, Sprachen und ihre Evolution.* Universitaetsverlag Regensburg, 2005.

[16] D. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2), 2006.

[17] D.H. Huson and T. Kloepper. Beyond galled trees: Decomposition and computation of galled networks. *RECOMB 2007*, LNBI 4453, 2007.

[18] International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, 1999.

[19] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22(21), 2006.

[20] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. A new linear-time heuristic algorithm for computing the parsimony score of phylogenetic networks: Theoretical bounds and empirical performance. *Proceedings of the International Symposium on Bioinformatics Research and Applications (ISBRA). Lecture Notes in Bioinformatics*, 4463, 2007.

[21] P. Ladefoged et al. Ucla phonetics lab data. http://phonetics.ucla.edu/, 2006.

[22] P. Ladefoged and I. Maddieson. *The sounds of the world's languages.* Blackwell Publishers Ltd., 1996.

[23] R. Lass. *Historical Linguistics and Language Change.* Cambridge University Press, 1997.

[24] I. Maddieson. *Patterns of Sound.* Cambridge University Press, 1984.

[25] A. McMahon and R. McMahon. Finding families: Quantitative methods in language classification. *Transactions of the Philosophical Society*, 10(1), 2003.

[26] D. A. Morrison. Phylogenetic tree-building. *Int Journal of Parasitology*, 26(6), 1996.

[27] M. Nei and S. Kumar. *Molecular Evolution and Phylogenetics.* Oxford University Press, 2000.

[28] J. Nerbonne, W. Heeringa, and P. Kleiweg. Finding families: Quantitative methods in language classification. In D. Sankoff and J. Kruskal, editors, *Time Warps, String Edits, and Macromolecules*, chapter Edit Distance and Dialect Proximity. Addison Wesley Publications, 1999.

[29] D. Posada and K.A. Crandall. Intraspecific gene genealogies: trees grafting into networks. *TRENDS in Ecology and Evolution*, 16-1:37–46, 2001.

[30] D. Ringe, T. Warnow, and A. Taylor. Indo-European and Compuational Cladistics. *Transactions of the Philosophical Society*, 100(1), 2002.

[31] N. Ritt. *Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change*. Cambridge University Press, 2004.

[32] D.F. Robinson and L.R. Foulds. Comparing phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[33] N. Saitou and M. Nei. The neighbor–joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[34] A. Schleicher. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. H. Böhlau, 1861.

[35] J.A. Studier and K.J. Keppler. A note on the neighbor–joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5:729–731, 1988.

[36] M. Swadesh. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137, 1955.

A Appendix

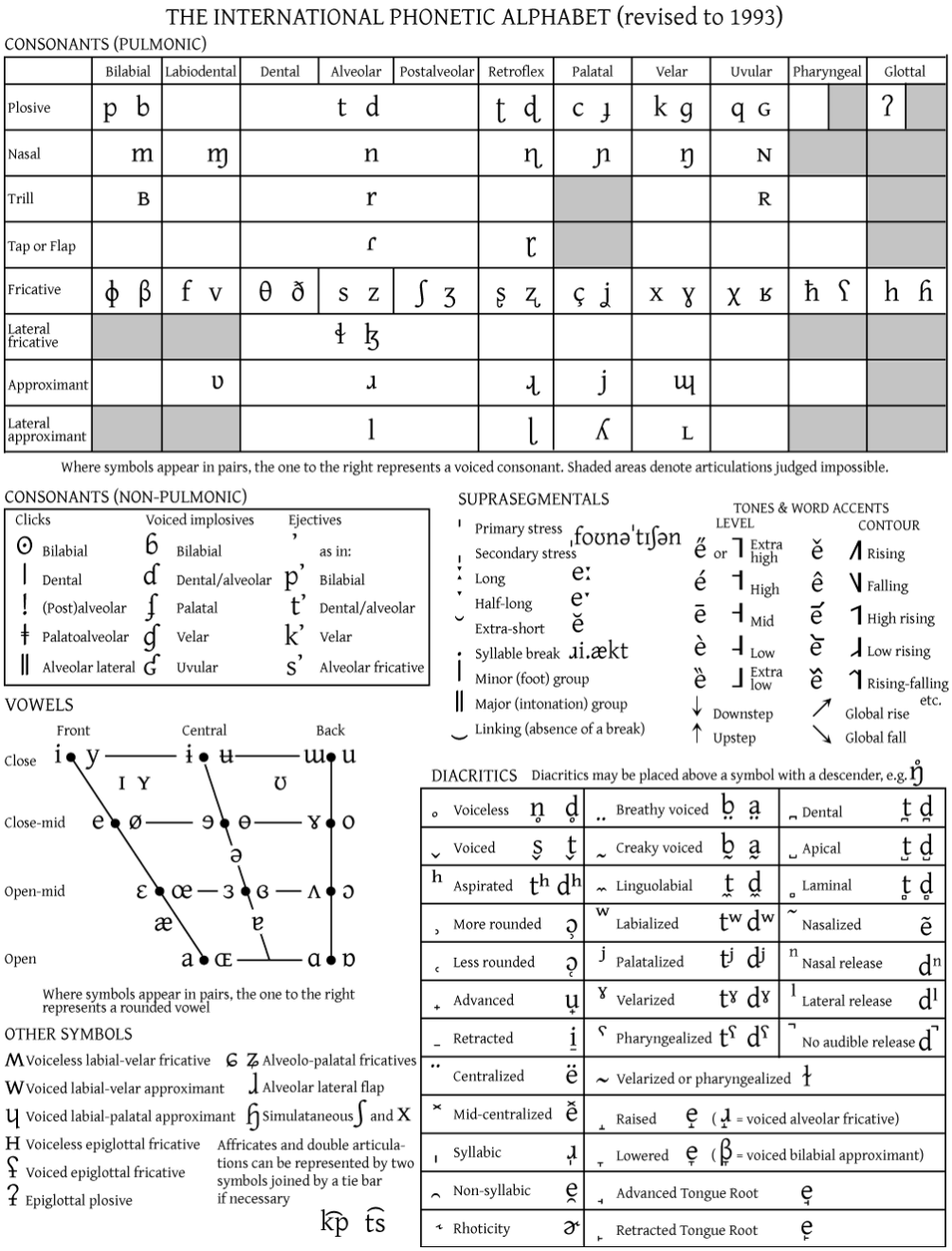

Figure A.1: The classification of sounds by the International Phonetic Assosiation (IPA).